Real-world Human Re-identification: Attributes and Beyond

Ryan David Conway Layne

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2015

Real-world Human Re-identification: Attributes and Beyond

Ryan David Conway Layne

Abstract

Surveillance systems capable of performing a diverse range of tasks that support human intelligence and analytical efforts are becoming widespread and crucial due to increasing threats upon national infrastructure and evolving business and governmental analytical requirements. Surveillance data can be critical for crime-prevention, forensic analysis, and counter-terrorism activities in both civilian and governmental agencies alike. However, visual surveillance data must currently be parsed by trained human operators and therefore any utility is offset by the inherent training and staffing costs as a result. The automated analysis of surveillance video is therefore of great scientific interest. One of the open problems within this area is that of reliably matching humans between disjoint surveillance camera views, termed *re-identification*. Automated re-identification facilitates human operational efficiency in the grouping of disparate and fragmented people observations through space and time into individual personal identities, a pre-requisite for higher-level surveillance tasks. However, due to the complex nature of real-world scenes and the highly variable nature of human appearance, reliably re-identifying people is non-trivial.

Most re-identification approaches developed so far rely on low-level visual feature matching approaches that aim to match human detections against a known gallery of potential matches. However, for many applications an initial detection of a human may be unavailable or a low-level feature representation may not be sufficiently invariant to photometric or geometric variability inherent between camera views. This thesis begins by proposing a "mid-level" human-semantic representation that exploits expert human knowledge of surveillance task execution to the task of re-identifying people in order to compute an attribute-based description of a human. It further shows how this attribute-based description is synergistic with low-level data-derived features to enhance re-identification accuracy and subsequently gain further performance benefits by employing a discriminatively learned distance metric. Finally, a novel "zero-shot" scenario is proposed in which a visual probe is unavailable but re-identification is still possible via a manually provided semantic attribute description. The approach is extensively evaluated using several public benchmark datasets.

One challenge in constructing an attribute-based and human-semantic representation is the requirement for extensive annotation. Mitigating this annotation cost in order to present a realistic and scalable re-identification system, is motivation for the second technical area of this thesis, where transfer-learning and data-mining are investigatedin two different approaches. Discriminative methods trade annotation cost for enhanced performance. Because discriminative person re-identification models operate between two camera views, annotation cost therefore scales quadratically on the number of cameras in the entire network. For practical re-identification, this

is an unreasonable expectation and prohibitively expensive. By leveraging flexible multi-source transfer of re-identification models, part of this cost may be alleviated. Specifically, it is possible to leverage prior re-identification models learned for a set of source-view pairs (domains), and flexibly combine those to obtain good re-identification performance for a given target-view pair with greatly reduced annotation requirements.

The volume of exhaustive annotation effort required for attribute-driven re-identification scales linearly on the number of cameras and attributes. Real-world operation of an attribute-enabled, distributed camera network would also require prohibitive quantities of annotation effort by human experts. This effort is completely avoided by taking a data-driven approach to attribute computation, by learning an effective associated representation by crawling large volumes of Internet data. By training on a larger and more diverse array of examples, this representation is more view-invariant and generalisable than attributes trained on conventional scales. These automatically discovered attributes are shown to provide a valuable representation that significantly improves re-identification performance. Moreover, a method to map them onto existing expert-annotated-ontologies is contributed.

In the final contribution of this thesis, the underlying assumptions about visual surveillance equipment and re-identification are challenged and the thesis motivates a novel research area using dynamic, mobile platforms. Such platforms violate the common assumption shared by most previous research, namely that surveillance devices are always stationary, relative to the observed scene. The most important new challenge discovered in this exciting area is that the unconstrained video is too challenging for traditional approaches to applying discriminative methods that rely on the explicit modelling of appearance translations when modelling view-pairs, or even a single view. A new dataset was collected by a remote-operated vehicle using control software developed to simulate a fully-autonomous re-identification unmanned aerial vehicle programmed to fly in proximity with humans until images of sufficient quality for re-identification are obtained. Variations of the standard re-identification model are investigated in an enhanced re-identification paradigm, and new challenges with this distinct form of re-identification are elucidated. Finally, conventional wisdom regarding re-identification in light of these observations is re-examined.

Submitted to the University of London in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

Queen Mary University of London

2015

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Parts of the thesis have been published previously:

Chapter 3

- R. Layne, T. M. Hospedales and S. Gong, *Attributes-based Re-identification*, In Gong, Cristani, Yan and Loy (Eds.), *Person Re-identification*, Springer, 2014.
- R. Layne, T. M. Hospedales and S. Gong, *Person Re-Identification by Attributes*, British Machine Vision Conference, Surrey, England, 2012.
- R. Layne, T. M. Hospedales and S. Gong, *Towards Person Identification and Re-Identification With Attributes*, Workshop on Re-identification (REID), European Conference on Computer Vision, Florence, Italy, 2012.

Chapter 4

- R. Layne, T. M. Hospedales and S. Gong, *Re-identification: Hunting Attributes in the Wild*, British Machine Vision Conference, Nottingham, England, 2014.

Chapter 5

- R. Layne, T. M. Hospedales and S. Gong, *Domain Transfer for Person Re-identification*, In Proc. ACM International Conference on Multimedia, Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS), Barcelona, Spain, 2013.

Chapter 6

- R. Layne, T. M. Hospedales and S. Gong, *Investigating Open-World Person Re-identification Using a Drone*, Workshop on Visual Surveillance and Re-identification, European Conference on Computer Vision, Switzerland, 2014

Ryan David Conway Layne

September 2014

Acknowledgements

Of all the sections in the following thesis, this one was the most difficult for which to find the words. I must first express my gratitude to my supervisors Shaogang Gong and Tao Xiang for their advice and relentless enthusiasm for our field as well as, of course, for providing so many of us with the opportunity to begin to contribute to it as their students. I am especially grateful to Timothy Hospedales for his enduring patience during our collaborative efforts, as well as his continued guidance, criticism, insight, and many valuable discussions.

I wish to thank both the academic staff and students from across our school of EECS – with whom it has been a pleasure to work alongside, and whose conversation I have enjoyed. In particular, Chris Russell, Tom Haines, Colin Powell, Chrisantha Fernando, Miles Hansard, Ravi Garg, Yanwei Fu, Parthipan Siva, Howard Wilson, Ho Huen, Linda Hogg and Visnja Curic, I thank for their invaluable advice, friendship, and counsel over the last three years. Also Lukasz Zalewski, Tom King, Tim Kay, Keith Bancroft and Peter Childs from the Systems team deserve special thanks for their patience with my crazy requests and for keeping my code running – if/when I got it to run in the first place.

I would be *highly* remiss if I did not also thank my MSc supervisor and friend, Mark Bishop, who inspired me to eschew random walks in favour of climbing this hill – often-times a steep one – toward a doctorate.

Finally, thanks to my parents, family, and friends for their unwavering support and for not being too cross at me when I've disappeared off the grid for weeks at a time.

– R

Contents

List of Figures

List of Tables

List of Acronyms

AOI Area of Interest

AUC Area Under Curve

BOW Bag of Words

CCTV Closed-Circuit Television

CMC Cumulative Match Characteristic Curve

CMOS Complementary metal-oxide semiconductor

CUHK Chinese University of Hong Kong (dataset)

DML Distance Metric Learning

ELF Ensemble of Localised Features

EOI Event of Interest

ER Expected Rank

FUSIA FUSed Internet Attributes

FV Fisher Vector

GRID QMUL underGround Re-IDentification Dataset

HUD Heads-Up-Display

IA Internet Attributes

MI Mutual Information

MRP Mobile Re-identification Platform

MSP Mobile Surveillance Platform

nAUC Normalised Area Under Curve

NN Nearest-Neighbour (distance)

OAR Optimised Attribute (based) Re-identification

PCA Principle Component Analysis

PRID Person Re-ID Dataset

PSD Positive Semi-Definite

RCCA Regularised Canonical Correlation Analysis

ROI Region of Interest

SAD Sum of Absolute Differences

SVM Support Vector Machine

UAV Unmanned Aerial Vehicle (Drone)

VIPeR Viewpoint Invariant Pedestrian Recognition (dataset)

ZSL Zero-Shot Learning

ZSR Zero-Shot Re-identification

Chapter 1

Introduction

"And set a watcher upon her, great and strong Argos, who with four eyes looks every way. And the goddess stirred in him unwearying strength: sleep never fell upon his eyes; but he kept sure watch always."

– Hesiod's "The Aegimius"

In the past decade there have been many technical innovations and advancements in the use of visual sensing technology, in large part due to increasingly cheap and powerful computer equipment. Indeed, computers and cameras can now be found in everyday appliances like mobile phones and bathroom scales, "smart" advertising billboards in the high-street and even worn upon our person or integrated into other personal accessories such as watches and eye-wear. Today most people own at least one hand-held device capable of communication and accessing information from the Internet. Computers have become adjunct to human cognition to the extent that many people would struggle without the ability to use a computational device of some kind, either extending their natural abilities and work efficiency or facilitating personal recreational activities. In today's society, the ubiquity of modern computing technology has infiltrated and become integrated into practically every aspect of our lives, and has become a vital tool that has improved our efficiency at performing many tasks. Computers are capable of performing simplistic tasks faster and more tirelessly than humans themselves, but this is not sufficient for the performance of higher-level real-world visual tasks that humans are able to do effortlessly such as identifying a friend in the street or recognising a co-worker's absence from a meeting.

These invariably require more advanced operations that cannot so easily be defined. Many daily activities humans perform involve multiple senses, amongst which the visual sense is prime. The visual sensing ability enables humans to distinguish salient objects and entities of importance in their immediate environment such as food, vehicles, animals and each other, as well as affording the ability to navigate complex and cluttered environments and to read, write and communicate. It facilitates these tasks even in the absence of other senses such as touch, smell, or hearing. Visual sense provides a constant stream of important and egocentric information, fulfilling the need for higher-level contextual cues that enable humans to take actions based on mixtures of often complex and nuanced visual observations. The benefit of replicating similar human visual functionality via artificial means is therefore predetermined and fundamentally important for a broad variety of traditional application areas such as robotics, industrial automation, or navigation which would benefit greatly from heightened abilities in these areas. New application domains continue to emerge as technology and human needs evolve, however. One such area is that of visual surveillance, important since it bears the potential to help prevent crime and crucially provides useful intelligence that may assist government agencies in reducing terrorist threats toward critical infrastructure and against the safety of citizens. Whilst human visual surveillance has been employed for many decades if not millennia, technological surveillance is still unable to replace the need for human insight. A human can recognise someone he or she has seen before despite poor lighting, an incomplete image and from irrespective of media type such as photographs, video or drawings. The act of recognition – or re-identification – is simple, but paves the way for more complex and useful downstream tasks and underpins the very essence of intelligence gathering and human visual analysis. Shortening the gap between human performance and machine-learning-driven, algorithmic performance at similar visual surveillance tasks is therefore highly vital.

1.1 Automated Visual Surveillance

The most popular and common realisation of visual surveillance technology in current use by human operators can already be found in every major city worldwide, making it the obvious candidate for providing surveillance data for automated systems to exploit. Closed-circuit television, (or CCTV) cameras, such as those normally seen affixed to ceilings or the sides of buildings, are the most commonly observable means of obtaining surveillance data, and in the past forty

years, CCTV has changed little in function or form excepting improved resolution and the incremental upgrade of CCTV cameras – so-called pan-tilt-zoom cameras – capable of interactive re-orientation by the operator in order to provide enhanced real-time surveillance coverage. The availability of cheaper and better surveillance technology such as CCTV, combined with escalating worldwide security challenges has encouraged governments to focus on deploying more and more preventative surveillance equipment [60, 5]. This has resulted in a widespread proliferation of video surveillance equipment for intelligence gathering, municipal crime and terrorism prevention, and other monitoring purposes such as for health and safety or business analytics [60]. The benefits of deploying surveillance in these areas are numerous; principally, events involving humans can be passively and unobtrusively recorded from a distance, often within large public spaces and over long periods of time. The recorded data are thus available if the need for detailed record arises in the future, but in practice storage is not infinitely available nor cheap and the effort and expense involved in having human operatives analyse data "just in case" is impractical. This poses a significant challenge to government agencies who wish to benefit from dense human-level intelligence, insight and description of surveillance data – which is more easily stored than retaining the video data itself – but cost-effectively and without the need for one human operator per camera.

1.1.1 Surveillance Technology and Operation

CCTV is most commonly deployed on a permanent or long-term basis in public spaces at fixed angle and elevation, and directed towards areas-of-interest (AOI). Primarily the motivation is to cover critical infrastructure or civilian crime and terrorism hot-spots such as transit hubs, public transport vehicles, or shopping arcades; wherever the risk of criminal or terrorist activity is expected to be either frequent or sustained [90].

Private entities may also deploy CCTV camera networks according to their own requirements, such as placing recording entrance and egress points, service areas in which employees may intersect with the general public, or high-value areas such as vaults or warehouses containing stock and equipment. Regardless of whether the operating entity is private industry or public sector, CCTV camera data are usually routed to a single location, normally a centralised operations room such as that depicted in Figure 1.1 on page 25 where one or more trained human operators will be employed.

Because not all CCTV cameras in the United Kingdom are centrally owned it is difficult to

quantify exactly how many are in use and for which purposes, however various reports indicate an increase from just approximately 100 cameras across three town centers in 1990, to 5,238 cameras across 167 towns in 1997 [5], approximately 2 million in 2006 [181] rising in 2014 to between 4.9 and 5.9 million CCTV cameras in the entire United Kingdom [1, 132]. Of the cameras estimated to be in use in 2014, it is further estimated 70,000 to 84,000 cameras are available for immediate use by government agencies [1] with a high proportion active in London and the other major cities [60]. The cameras are used in the execution of various surveillance tasks by human operators, most notably for:

1. Tracking target individuals through a distributed camera network
2. Identifying target individuals from prior "watchlists"
3. Identifying suspicious behaviour, objects, or vehicles
4. Identifying accidents or emergencies
5. Monitoring human or vehicular traffic patterns and flow
6. Monitoring crowd behaviour

However, the use of human operators for CCTV monitoring is costly and inefficient. Operators must be trained in order to make use of CCTV footage effectively, since there is a significant performance deficit between untrained and trained operators [132, 180]. In addition, standard CCTV control room practices have been repeatedly shown to be inefficient [181, 89, 90, 160, 63, 62, 79, 35], but remain mostly unchanged since the 1970s in terms of practices [1, 90, 35] and indeed control room configuration and structure [1, 35] (See Figure 1.1 on the next page for a visual comparison of control rooms in the 1970s and in recent years).

CCTV operator efficiency has been investigated regularly since the 1970s with slightly more emphasis on the psychological motivations of operators rather than their capability at specific tasks or the overall control-room paradigm. Early research by Tickner *et al*. indicated that operators may only monitor less than 10 cameras before their performance was significantly impaired at standard detection tasks, due to perceptual overload [168]. Gill *et al*. note that as of 2005, the operators they observed were responsible for up to 90 cameras at a time [63]. Keval and Sasse, and in separate work Smith, report the average work shift of a modern CCTV operator was now 12-hours [90, 160] – whereas a reasonable expectation of alertness and attention to a

Figure 1.1: The evolution of modern CCTV control rooms; (left) Circa 1970, a control center in Munich, Germany; (right) Image of a contemporary CCTV control room from Sedgemoor Council, England. Despite the large amount of visual data available, only a small number of operators will be monitoring in real-time, and despite advances in storeage technology, recorded video is often deleted after a month.

retrieval or search task is just 50 minutes before a rest period is required [175]. Smith describes the day-to-day effect of long-term CCTV operation on human operators under standard working conditions, listing multiple deficits; task-avoidance behaviourisms such as smoking cigarettes, socialising, and abuse of CCTV systems for personal amusement, as well as feeling undervalued and immured by their work [160]. Both Smith, and Dadashi *et al*. conclude that these factors significantly undermine any potential CCTV surveillance effectiveness. These underlying issues clearly undermine operator vigilance and workload and irrespectively even motivated and highly trained operators must take breaks, eat, and communicate during which time their attention is not focused on the task of surveillance. Furthermore, the transfer of tacit operational knowledge between operators during training of new recruits is inherently lossy, resulting in lead-times before each operator can achieve full operational performance levels.

Although these studies are sometimes isolated or conducted at small scale, the factors highlighted are of clear relevance and generally motivate the need to assist human operators with day-to-day surveillance tasks using automated technology.

1.2 Mobile Surveillance

Recently, commodity products such as smart phones, passenger vehicles, remote-operated vehicles and even eye-wear are capable of recording quality video. This has given rise to a potential new modality of surveillance source footage. One particular such alternative source is underpinned by the escalating use of remote-operated vehicles, or unmanned aerial vehicles, colloqui-

ally referred to as UAVs or "drones" [34, 134]. UAVs have become widely affordable and are now increasingly available to civilians rather than solely to government agencies – much like CCTV rapidly became widespread from the 1970s onwards. So-called "off-the-shelf" UAVs are commonly equipped with visual sensors of rival or better quality to contemporary CCTV cameras. Although not currently as prevalent in surveillance applications as CCTV cameras, the use of diverse devices is becoming increasingly common and may afford additional capabilities which have yet to be considered fully for automated surveillance tasks [34].

While the *de facto* sources from which most video surveillance is derived are statically-placed CCTV cameras, the technology has limitations. The advent of affordable and widely integrated visual sensing equipment into everyday artefacts such as phones, tablets and even vehicles and clothing, provides further opportunities for the exploitation of video data for surveillance purposes. The range of suitable devices is very broad, but all devices can either record or stream video data and are qualitatively more flexible for surveillance due to being mobile. In this thesis they are broadly designated *mobile surveillance platforms*, or MSPs. Table 1.1 on the facing page provides comparisons between the principle differences of CCTV and MSPs. Non-exhaustively, some of the primary benefits of using MSPs for surveillance instead of, or to augment static CCTV are:

1.2.1 Rapid Deployment

MSPs may be as small as a personal mobile phone or mounted on a remote-operated or human-driven vehicle, or integrated into personal clothing, permitting a major change in the way surveillance tasks can be conducted. In traditional surveillance networks, each camera is permanently placed and is immobile. Additional cameras may be added at the cost of additional cabling for power and data transfer, connected to the existing systems. Conversely, additional MSPs can be deployed far more quickly than static CCTV cameras, albeit with the disadvantage of reduced operational durations. Despite this trade-off being able to quickly deploy an entire surveillance network means that surveillance can be performed *ad-hoc* and in public spaces where current CCTV coverage is non-existent or otherwise poor.

1.2.2 Mobility

Perhaps more crucially, such networks of MSPs are intrinsically dynamic in nature and can reposition, re-orient, or follow as required by circumstance, subject only to constraints on power,

	Degrees of Freedom	Operational Duration	Data transmission	Operating Mode
CCTV (Static)	0	Years	Cabled, some wireless	Passive
CCTV (PTZ)	2, (pan, tilt)	Years	Cabled, some wireless	Passive, Autonomous, Interactive
MRP	6, (position, yaw, pitch, roll)	Up to 72 hours	Wireless	Autonomous, Interactive

Table 1.1: Key differences between types of surveillance technology, standard closed-circuit television (CCTV), CCTV with pan-tilt-zoom capability (PTZ) and mobile re-identification platforms (MRPs), such as UAVs or portable cameras.

communication reception and mode of transport (such as whether it is carried by a human or mounted on a UAV). This flexibility is of great utility in situations where surveillance must be continuously maintained across distance, repositioned or reorientated rapidly. Lastly, where dynamic scene clutter or occlusion prevents adequate surveillance of a particular target, the ability to move to a new relative viewing angle could be critical. Mobility also introduces a detrimental factor in addition to the standard challenges with visual sensing equipment, as translational and relative motion can introduce further complications such as "motion blur".

1.2.3 Autonomy

Complete autonomy for the smaller UAVs is currently limited although it is common to find subsets of obstacle-avoidance, round-trip navigation and rudimentary visual-sensing in popular commercial offerings [34], this is mostly due to the limitations of on-board processing, load-bearing and power supply. Larger UAVs possess richer capabilities, particularly military-operated UAVs, but are unable to manoeuvre at distances close enough to be useful for classic surveillance tasks and cost significantly more [54] (Figure 1.2 on the next page illustrates the most salient viewpoint and modality differences between common surveillance data sources). Nevertheless, complete autonomy is a useful scientific area for surveillance tasks, permitting the operation of UAVs and other self-propelling visual sensors to patrol areas or perhaps dynamically "chase" targets of interest, or otherwise perform specific tasks attentionally as required like moving closer to a suspicious person detection in order to get a better reading, or otherwise optimise position for some task [54, 57, 177, 127].

1.3 Human Re-identification

The general surveillance scenario is that cameras are placed and then analysed in real-time or after an incident and a variety of tasks must be performed in order to obtain good intelligence.

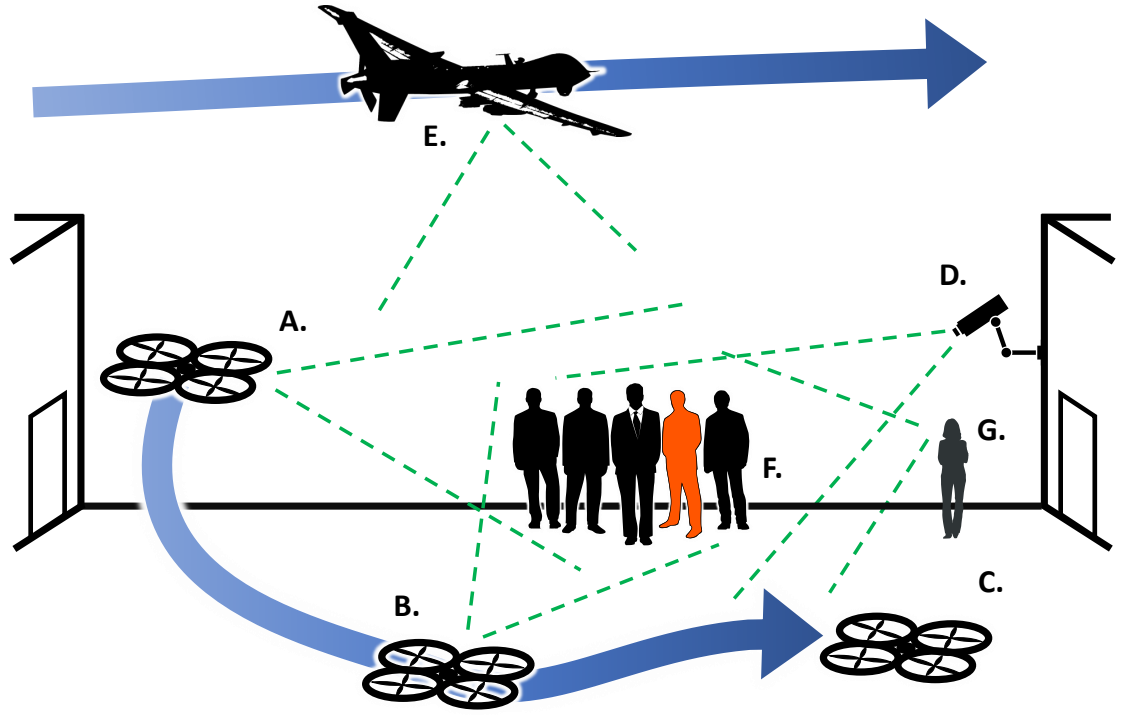


Figure 1.2: Comparing viewpoint variations, dependent on surveillance video source, for the standard re-identification problem; re-identifying a specific target (red silhouette) from other people (F). Blue arrows denote typical approach vectors, field of view is illustrated with green lines. Quadrocopter-type UAVs possess high mobility and can be operated at variable range (A-C), thus surveillance applications from this source must be robust to extreme viewpoint variability. Larger UAVs (E) operating at higher altitudes are somewhat more constrained but still highly variable. In contrast, closed-circuit television (CCTV) provides an immobile view (D). Finally, human-portable devices permit closest-range observation (G).

Where these tasks are aimed at providing intelligence on human entities however, there is a single task that underpins most others, aside from detecting humans themselves: reconciling the detections into entities, or *human re-identification*.

Human re-identification, refers to the task of recognising a particular individual in diverse scenes obtained from non-overlapping cameras (Figure 1.3 on the following page). Specifically, for surveillance applications performed over space and time, re-identification is the fundamental task that permits an individual transiting from one view to be differentiated from numerous possible targets and matched in one or more other views at different locations and times. This task therefore underpins and forms the foundation of a large number of crucial surveillance tasks such as longer-term multi-camera tracking and forensic search, criminal investigations and intelligence-gathering. Success at the re-identification task therefore paves the way for richer surveillance data and applications and not just retracing the steps of a particular individual; aggregating large volumes of individual observations reconciled by person identity can also grant important insight into crowd-movement and facilitate planning operations, seasonal appearance and behavioural model formation and anomaly detection tasks.

In conventional real-world surveillance scenes, there are too many unconstrained factors such as lighting, distance from the camera to the person and person pose, to rely upon higher-level biometry such as face recognition as intuition suggests we might. Indeed, if faces are detectable at all they will only rarely be detectable at sufficient resolutions to be useful. Instead, holistic appearance models are usually constructed taking into account the entire appearance of an individual – clothing being the predominant cue, as well as hair or skin colour and carried objects. However, this approach is inherently weak and does not generalise perfectly. For instance, darker clothing is predominant in winter which limits the discriminativeness of this kind of appearance model. Furthermore, there are no guarantees that an individual's relative orientation to the camera will be the same in each camera view, the result of which being that in order to re-identify a person under these conditions a system must be able to match the front of a person to the rear view of the same person and disambiguate between other, more similar but incorrect people. More formally, intra-class variability is frequently going to be significantly larger than inter-class variability across cameras.

Setting	Camera Pairs	Match	Person Count	View-specific	Multi-shot	Evaluation
Singleshot [47, 186, 94, 6]	1+	$N:N$	Known	Yes	No	Rank 1, CMC
Multishot [47, 91]	1+	$N:N$	Known	Yes	Grouped	Rank 1, CMC

Table 1.2: There is little variation amongst standard re-identification problem variants besides having one "shot" of each person per view, or more shots of each person per view. Match: $N:N$ reflects closed world one-to-one mapping among N people in view 1 : view 2. This table is expanded later in Table 6.1 on page 151, Chapter 6, where additional formulations are posited.

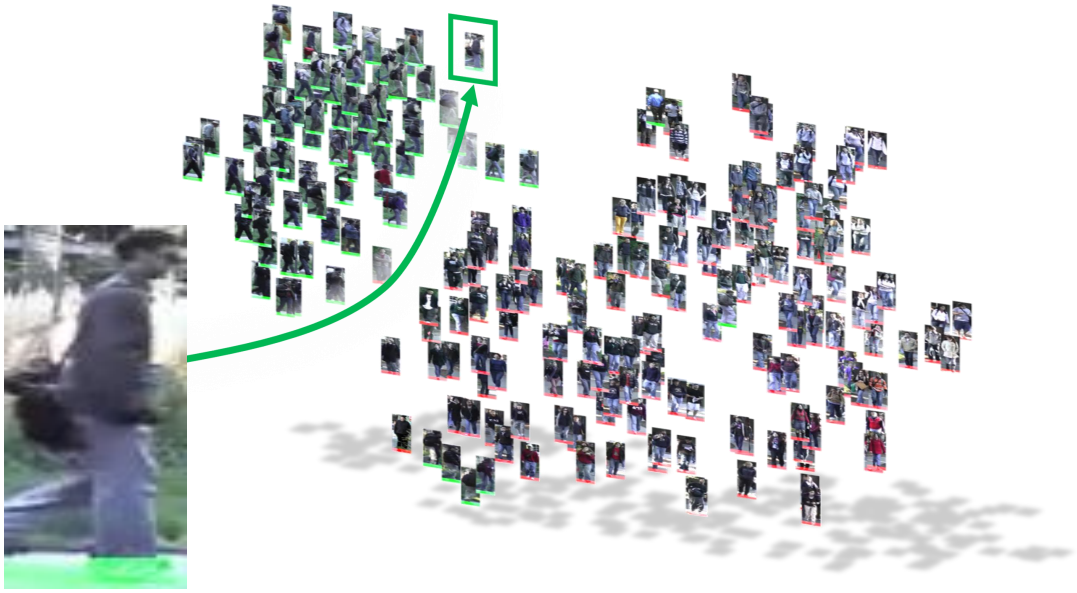


Figure 1.3: An illustration of the re-identification problem. Given a "probe" image of a person observed somewhere in the surveillance network (left), a subset of all possible matches to search form a "gallery" (right), with the goal being to correctly identify the image containing the person shown in the probe (correct match highlighted by a green box).

1.4 Challenges and Motivation

Conceptually, the re-identification task could appear to be a simple retrieval task between individual camera views; indeed, some surveillance tasks such as querying a camera network for people wearing "red shirts" are very close to this definition. However, the key challenge for re-identification is not to simply locate people with similar appearance, it is to distinguish within those people to locate the same *identity* from all the others, with appearance being a potentially confounding or assistive factor. Since most CCTV video footage records humans at stand-off range, reliable biometry such as iris or face recognition is generally not possible without specialist equipment. The majority of re-identification methods utilise the person's overall appearance

instead, but this creates several severe challenges [65].

1.4.1 Viewpoint and Appearance Variation

By definition, re-identification is normally performed between two or more unconstrained but fixed views, meaning that observations of people from each camera will be influenced by different visual factors. This is due to the different positions, model, calibration, positional elevations and angles of the video sources used, as well as factors inherent to each location such as scene clutter or lighting change caused by environmental factors (see Figure 1.5 on page 37 for some comparative surveillance video frames, and Figure 1.4 on the following page for examples of the variability of person appearance). Each of these factors may change independently and may be variably constrained.

1.4.2 Person Appearance

A person's appearance can vary dramatically depending on their underlying build, taste in clothes, pose, where they are observed, and multitudes of other factors. Generally, apparel breaks down into five categories; (i) upper-body clothing (ii) lower-body clothing, (iii) full-body clothing and (iv) head and footwear, and (v) accessories and carried items. Different clothing types possess differing degrees of variability in terms of shape or state, for example long skirts, long hair and scarves might vary dramatically in appearance depending on whether they are subject to wind whereas cycling shorts or t-shirts may not due to being closer fitting, and jackets or coats might be open or closed thus revealing lower layers of clothing that may not be visible from the side or rear. Carried items can be useful discriminatory appearance cues for some surveillance tasks but may be observed in many different configurations and relative positions to the person carrying them, as well as left unattended elsewhere. Accessories such as bracelets and necklaces may not be reliably visible at all except at very close range from the camera, but may be inferred occasionally from subtle clues such as specular reflections [132].

Personal apparel can additionally consist of different colours, patterns or logos per item, both front and back; as well as different trim, detail, and features all of which may be coloured and textured uniquely. A further complication is that colour can be affected by environmental factors such as local lighting condition; one example of this case might be observing a person wearing a seemingly black top, when in fact the top is red and illuminated by a blue light.

Appearance is therefore a wealth of potentially discriminatory information for surveillance

Figure 1.4: Appearance factors critical to matching pairs of person detections from the VIPeR dataset [67]. Clockwise from top; consistent texture, consistent colour, texture concealed from side view due to jacket, texture concealed due to self-occlusion and lighting condition change, texture change due to apparel deformation and self-occlusion, different logo on front and rear view.

purposes, but this information may be terminally hard to utilise effectively by automated systems due to the extremely high number of possible combinations and permutations in unconstrained scenes, particularly when relative viewing angle is not constrained. Figure 1.4 illustrates the major forms of appearance variation as seen in one dataset [67], formally introduced in Chapter 3.

1.4.3 Stand-off Range, Enrolment and Biometry

The intuitive method of performing re-identification is to find invariant cues that guarantee identity can be determined with little or no ambiguity, however such approaches are problematic for the standard visual surveillance setting. Biometrics refers to such cues, unique to the individual, that can be both quantitatively determined and remain robust over time and location. Some examples of biometrics include fingerprints, retinal patterns, gait, typing patterns, and the written signature. Whilst unequivocally useful cues for re-identification in theory, in practice except for gait, these biometrics all require at least one of: (i) inconvenient and invasive enrolment

procedures, (ii) active participation, (iii) and specialist sensing equipment. This makes them potentially viable for closed-world environments such as secure buildings, but not for public space surveillance with existing visual surveillance sensors already *in situ*.

Common types of enrolment include having high-resolution photographs taken at many different angles, fingerprint registration, or typing tests; which all require the active participation of the subject before any surveillance task is possible. Even though such methods may take between just minutes to an hour to complete, it is not feasible or desirable to employ them for surveillance purposes at busy transit hubs and impractical do so for other public spaces with no distinct bottlenecks; moreover, such processes would be inconvenient and actively rejected by the general public on grounds of privacy. Therefore, any surveillance of public spaces must be capable of success relying only on passively collected data and not rely on elaborate enrolment schemes. This means that whilst surveillance equipment may be visible, and emplaced in such a position as to maximise the chances of observing useful parts of a scene, that observation is carried out from a distance and without disrupting the normal activities of the people transiting the area.

So-called "soft"-biometrics are a compromise between enrolment, requiring participation and discriminativeness. They provide identifying cues that are not univocal to each individual person but which may be sufficient in combination even amongst a large gallery of potential "hits" to search within. Commonly soft-biometrics such as height, tattoos, facial hair, scars, gait, body/limb proportion and hair and eye colour are used in this manner by human specialists. For common surveillance scenarios such identifying cues are often impractical depending on the specific scene configuration, *i.e*. due to the resolution of the camera and occlusion due to apparel, and the distance of the subject from the camera.

The distance between the camera and the human, called stand-off range, is an important factor. Potentially this distance is constrained by the physical configuration of the public space (walls, fences, foliage, paths, doors) as well as other minutiae such individual behaviours; for example, a camera viewing a busy transit hub will observe the majority of people following standard routes with little deviation over long periods of time, but consider the rarer cases where there is a new advertisement or poster of information on a wall near a less-travelled area. The information available at that specific location may only be of interest to a small percentage of people overall or be only temporary, but could well result in a shift in the distribution of observations at a particular distance and orientation from the camera.

Generally, distances that are too great from the camera will result in low-resolution person detections that may not be directly useful for re-identification, for example because a distinctive visual feature such as a logo or texture pattern may not be visible or the overall appearance is not discernible from the amount of information available.

For practical applications, some of these factors may be alleviated by giving consideration to to said factors during system design. As an example of this, consider the problem of people appearing too far away from the camera. In order to mitigate this problem, one may ensure surveillance cameras are placed at reasonable distances from the most frequently travelled areas of a scene, whilst ensuring the detected people are observable at useful scales (a naive form of calibration). However, systems installed without automated surveillance specifically in mind may not be relocatable, or relocatable cheaply.

1.4.4 Intra View and Inter-View Variability

For surveillance tasks to be performed, data are taken from cameras occupying a unique location in space and providing a similarly unique viewpoint of a given scene. For the majority of cases and models of camera, this "view" of the world is fixed; the view will not rotate or otherwise change position relative to the scene being observed, thus for re-identification performed between such views to be successful, a critical part of the re-identification task concerns itself with matching people between these different views. One aspect of achieving this, is accounting for view-specific considerations.

The clutter and topography inherent to each specific and unique location will heavily influence the transitory paths humans select when crossing through the location, as well as the particular layout and current crowdedness of the location. These factors directly influence and constrain the set of all angles and frequency at which humans are observable by a single static camera. For example, a human face is much more recognisable from the front than from the side, and not at all visible from the rear; likewise some apparel might feature distinct designs or logos on the front or back which might be useful in determining human identity, but which are not visible all the time nor from every possible angle.

The appearance of a person can therefore radically alter depending on how it is viewed, especially when only a single observation of the person is available from each camera and other context is scarce. In these cases it could be easy to erroneously re-identify a person by falsely matching appearance cues between people who resemble each other more closely from some

angles than the true match resembles itself, solely due to circumstantial and unpropitious poses.

In addition, although in most current research, tasks are performed between two or more camera views, there exist multiple special cases where this assumption does not hold (particularly in real-world applications), for example when a surveillance target re-enters the same view or the view itself is non-stationary.

1.4.5 Within-view Ambiguity

Standard fixed camera re-identification assumes a set number of views between which to perform re-identification. That is, the standard setting is typically defined across a pair of camera views, and within-camera tracking is typically assumed to fully disambiguate detections within-view. For some applications, 'within camera' re-identification is necessary due to the lack of annotation effort or tracking capability, particularly evident in a scenario where the task is to perform real-time re-identification and from dynamic (non-stationary) views.

This is considerably non-trivial for the cameras with positional and orientational mobility, since this opens the possibility that even stationary people can enter and exit the view area solely due to the self-motion of the view.

1.4.6 Other Issues

The second challenge arises where viewpoint continually varies, perhaps because the camera is mounted on a mobile vehicle or is hand-held, rather than the conventional fixed position CCTV camera scenario. This is significant because for the most part, re-identification research follows human expert practices and training. For the fixed-camera case, an operator becomes familiar with a pair of camera views through repeated analysis. With a single continuously varying camera view undergoing constant changes in range, lighting, motion and position, a different set of skills and experience is required in order to compensate.

Most existing re-identification studies make the simplifying assumption of closed-world conditions. That is that there is a one-to-one set match, where everyone in the first camera re-appears in the second camera. No one disappears, and no extra people appear. Although convenient for modelling and benchmarking purposes, this is clearly an extremely strong assumption to make for practical applications. Given the mobile nature of some camera views, closed-world is clearly an inappropriate assumption – meaning that re-identification becomes significantly more ambiguous.

1.4.7 Supervised Learning, Annotation and Data Availability

A central limitation inherent to supervised learning approaches to automated re-identification that exploit human labelling in order to "learn" a more discriminative matching method is that such methods are more suited to closed-world benchmark scenarios rather than realistic open-world scenarios. The reason for this is that they require many pairs of person images annotated as having the same identity or not, *for each pair of cameras* between which the system is expected to operate. This is reasonable for synthetic studies and benchmark dataset volumes that are already exhaustively annotated for identity, however it is highly impractical for real-world use where many more cameras may be present in the network and where *each pair of cameras would require exhaustive annotation*, making deploying such a network laborious as well as prohibitively expensive. Ideally, one would wish to deploy a re-identification system between all camera pairs with minimal annotation and what a system learns from annotations on one camera pair should be exploited efficiently and effectively by the others without requiring much further effort.

Aside from being a crucial factor in determining the tractability of training discriminative matching classifiers, annotation cost is also a deciding factor for representation engineering. It is generally the intuition that improving a representation somehow, as well as selecting the most suitable discriminative learning method in tandem, is worthwhile in the sense that both representation and learner are linked in a synergistic manner and improving one improves the other with the reverse also being true. One may expect that better representations make the task of discriminative learning easier or more efficient, which in turn can provide better accuracy, generalisability or better performance from the learnt model. However another crucial factor to consider is the type and volume of human annotation work required in each of the previously mentioned cases. Representation learning based methods provide a means of constructing powerful feature representations, and do so at the cost and reliance on exhaustive human annotation. In the case of attribute learning, annotations must be supplied for each attribute, and on each dataset. For supervised re-identification models, pairs of person detections must be annotated with which to construct a binary classifier capable of determining whether a given tuple of images are of the same person. The annotation cost of this inter-camera case is therefore dependent on the number of possible camera *pairs* within the whole network. Therefore, for surveillance task representations using machine-learning methods, the volume of annotation required to train a sufficiently general discriminative model for real-world deployment in unconstrained environments is likely

	Ch 3: Attribute Learning	Ch 4: Internet Attributes	Ch 5:Transfer Learning	Ch 6: UAV Re-identification
Identity Labels	Needed for weight learning phase ✓	Needed for weight learning phase ✓	Use fewer ID labels ↓	Unavailable in real-time ×
Attribute Labels	Needed for Attribute Learning ✓	Discovered Automatically AUTO	Not Required ×	Not Required ×

Table 1.3: A tabular description of distinct types of annotation (identity or attribute) and annotation treatment (needed, not needed, or not used).

to be intractably costly. Table 1.3 describes the technical chapters to follow and their relationship with each type of annotation.

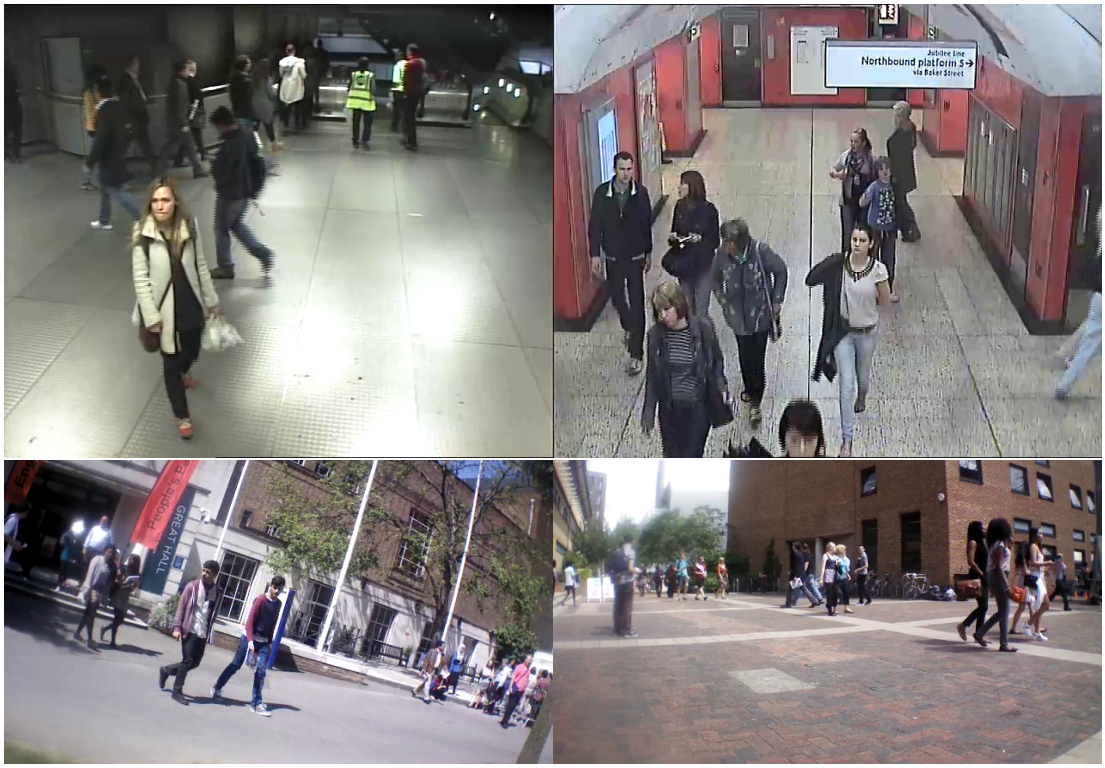


Figure 1.5: Demonstrating the visual differences between types of surveillance cameras, particularly between static-mounted CCTV cameras (top row), and mobile re-identification platforms (MRPs) such as low-altitude UAVs (bottom). All cameras feature scene clutter, target occlusion, and lighting and pose variations. MRPs offer further challenges due to their mobility which results in fully unconstrained pose variation versus the coarse constraints of static camera placement.

1.5 Thesis Overview

1.5.1 Robust Representations

Earlier in this thesis, the importance of overcoming camera-specific condition variations was highlighted, as well as the case that humans and computers have different abilities in terms of their current ability to semantically interpret video data, requiring specific treatment in order to conceptually interface human expertise to an automated video analysis system. Since human observations via multiple cameras from different locations can present significantly different appearances, the starting point of this thesis is to construct robust representations, ideally invariant to pose, background, lighting and occlusion, in order to facilitate subsequent re-identification.

Feature-centric approaches [47] suffer from the problem that it is extremely challenging to obtain features that are discriminative enough to distinguish people reliably, while simultaneously being invariant to all the practical covariates such as motion blur, clutter, view angle and pose change, lighting and occlusion. In contrast, learning approaches [77] make better use of a given set of features by discriminatively training models to maximise re-identification performance, for example metric learning [77] and support vector machines (SVM) [147, 6].

A mid-level semantic representation that is robust to the previously discussed challenges and also enables "querying" the surveillance network by description only. This permits re-identification even in the absence of a visually observed probe image and may be constructed via expert human guidance. Furthermore, a mapping function that infers the inter-attribute utility or usefulness is constructed for relatively low computational cost using Support Vector Machines (SVMs) [155] and standard optimisation methods [131]. This function is necessary since it is often hard to intuit *a priori* whether a visual attribute is at all tractable for discriminative classifiers to learn nor indeed whether it will be discriminative for identity, thus learning an inter-attribute weighting serves to reduce the noise contributed by weak classifiers whilst preserving the most useful attributes.

1.5.2 Reduce Annotation Cost: Gain Scalability

Mid-level semantic representation learning requires significant overheads of human annotation effort. In addition, discriminative modelling techniques commonly employed to achieve superior matching performance (for example, compared to nearest-neighbour matching) must be trained in a binary (same class versus different class) rather than a multi-class (person identity as indi-

vidual classes) setting. A central limitation of using such techniques in existing machine learning approaches is that for the re-identification problem, they are more suited to closed-world and less challenging benchmark problems than more realistic real-world scenarios with large numbers of cameras. Both representation learning and re-identification model learning require either pairs of annotated person images as being the same or different, *for each camera pair*, or individual person images annotated with one or more visual attributes. This is only tractable for benchmark datasets that are either already annotated by person identity, or can be exhaustively labelled for attributes by a diligent researcher. However, these same requirements are thoroughly impractical for systems that must scale for real-world use; it is extremely likely that the network will consist of too many cameras. This makes methods involving the training of a re-identification model for each camera pair, or the training of attribute classifiers for each attribute and each camera, prohibitively expensive.

Two broadly different approaches are explored to this end; (i) transferring previously-learned models to target domains using fewer human-annotated ground-truth label volume, and (ii) harnessing noisy Internet-sourced social media images and meta-data in order to construct bottom-up representations without the exhaustive annotation requirement of the previous work.

1.5.3 Transfer Learning

What a system learns from annotations of one camera pair should be exploited by another pair without requiring exhaustive annotation in the new pair. This is an issue in *transfer learning* [140, 45, 83]. Transfer learning is already important for many classical vision problems such as object recognition [151] with multiple classes or domains. However it is critically important for training re-identification models because the number of domains (camera pairs) can be *quadratic in the number of cameras*. Therefore, obtaining exhaustive training data for each domain is even more impractical than for conventional vision applications, thus transfer learning becomes critical. Despite this, no prior re-identification studies have addressed this issue. Our first approach toward alleviating the annotation cost employs the Multi-Kernel-Learning (MKL) approach from [46] to learn camera-pairwise non-linear decision boundaries from multiple source domains. These domains are subsequently projected onto a target-domain in order to improve learning for both sparse and even non-sparse training-data volume whilst avoiding so-called "negative transfer" (where transfer negatively impacts end performance rather than enhancing it) even if multiple source domains are irrelevant for the target.

1.5.4 Internet-driven Attributes

The first contribution of this thesis draws upon inspiration from the practices of human experts to learn an attribute-centric, low-dimensional feature representation that corresponds to semantic properties; but such top-down human-defined attribute approaches have some critical limitations: (i) They require costly attribute annotation of site-specific training data. This is significantly more laborious than the person-identity information used to train discriminative matching models. (ii) The top-down definition of attributes does not guarantee that they are visually computable by computer vision techniques given visual surveillance data. (iii) Due to the limited scalability of the annotation approach, the annotated data are likely to be too small scale to learn accurate and robust detectors for each attribute of interest.

The second approach addresses these issues by taking a very different data-driven [30, 126] approach to learning attributes for re-identification. In it, an automatically defined ontology is constructed from the bottom-up, as opposed to exploiting expert knowledge, and from it an effective associated representation is learned via the large-scale mining of noisy but abundant content on social photo sharing sites. Specifically, rather than asking an expert to define an ontology [101, 102, 103, 117, 153, 174], we discover it automatically by clustering photo tags and comment data. These clusters are used to train a bulk array of detectors using Linear Discriminant Analysis (LDA), resulting in a large number of visually detectable attributes (in contrast to expert defined ontologies, which while intuitive to experts, may require additional visual properties or otherwise may not be possible to detect reliably with current vision techniques). The greater volume and diversity of data used to train these automatically discovered attributes results in a more reliable and generalisable attribute representation than conventional attribute representation approaches on surveillance datasets can normally achieve.

1.5.5 Testing in the Open World

To ease model creation, evaluation and the establishment of benchmarks, most re-identification work is formalised as a closed-world set match between a single pair of specific cameras, given single observations of each person in each camera. As a result the typical evaluation metric is Rank 1 accuracy (the % of perfect gallery matches for each probe image), or the cumulative match characteristic (CMC) curve (the % of correct matches within the top N ranked matches, for varying N) [178]. In this thesis, this is referred to as the *standard re-identification problem*.

A very close variation is the *classic multi-shot re-identification problem*, which groups multiple observations (shots) by identity. Both the classic re-identification approaches assume a "watch-list" surveillance task or an inter-camera entity-association task and likewise tend to assume that for each of the N probe people, a true match exists in the gallery set of all possible matches. While a reasonable starting point for re-identification research, this scenario is artificially sterile and does not commonly arise in real-world re-identification applications. Table 1.2 on page 30 summarises these classical approaches to re-identification, which are extended in Table 6.1 on page 151, Chapter 6.

The majority of this thesis reflects the current state of art in person re-identification in terms of both the standard problem definition and set of assumptions employed in the exploration of it, however some of these assumptions do not apply in the conventional sense when considering an all-aspect re-identification system for use in the real world as well as with today's technological advances. Of those advances, this thesis primarily considers the introduction of UAVs and portable camera equipment for surveillance and attempts to reconcile this with existing contemporary re-identification assumptions and rationales. To do this, a small commodity remote-piloted flying vehicle was re-purposed and subsequently operated to perform surveillance tasks on a busy university campus across distributed locations and using a real-time person detector to cue the behaviour of the pilot to simulate a closed-loop, autonomous vehicle concerned with re-identifying people.

1.6 Thesis Contributions

The contributions of this thesis to human re-identification research are:

1. A re-identification-centric attribute representation, modular in the sense that additional mid-level semantic cues can be added and re-calibrated easily, and the final representation can be fused with additional information sources to in order to further improve re-identification and maximise early-rank or overall performance. Additionally, the representation is arbitrarily low-dimensional, an attractive property that facilitates tractable combinatorics, optimisation and distance metric learning. Finally, as the representation is readily human-interpretable, this permits *Zero-Shot Re-identification* (ZSR), a procedure where the visual probe is replaced with a manually constructed, human-defined probe vector [101, 102, 104]; this facility is crucial for real-world applications such as "per-

son searching" where an initial image is unavailable or a subset of individuals are to be retrieved given some shared appearance attributes.

2. We relax an important and practically unrealistic assumption, that there are exhaustive and readily available amounts of training data within each domain, by generalising recent ideas in discriminative-learning based re-identification [6] and SVM transfer learning [83]. Specifically, we consider re-identification based on binary-relation learning [6, 96], and show how to generalise this approach to achieve effective cross-domain learning by combining non-linear decision boundaries from source domains to create a more accurate target domain re-identification classifier. In this way we are able to improve on within-domain learning both for sparse and even non-sparse training data volumes. Moreover we show how to achieve this while systematically avoiding negative transfer, even when there are multiple irrelevant source domains.

3. A novel perspective on the re-identification challenge, driven by recent technological innovations in the fields of remote vehicle operation and the portability of visual sensing equipment as well as a global heightened need for surveillance coverage beyond static CCTV cameras. This part of the thesis makes four main contributions: (i) it presents a case for the pursuit and development of a new research area using mobile re-identification platforms (MRPs); (ii) it formalises three novel MRP-related variants on the classic re-identification scenario, as well as associated evaluation metrics for each; (iii) it describes the creation of the first known public dataset for MRP re-identification and establishes benchmarks for each of the identified tasks; (iv) finally it elucidates the unique challenges posed by MRP re-identification and discuss their implications for general re-identification research going forward.

1.7 Thesis Outline

The remaining chapters of this thesis are organised as follows:

- Chapter 2 is a review of contemporary research relevant to the main components of this thesis, including the extraction of higher-level representation from video surveillance data, learning to re-identify using discriminative models and low-level features from both video surveillance and Internet data, and the transfer of learned models to new data.

- Chapter 3 describes the process of learning a new, surveillance-specific attribute representation, and explores the benefits of this representation through subsequent re-identification experiments.

- Chapters 4 and 5 detail two distinct methods that both assist in the mitigation of annotation costs normally associated with the use of state of art discriminative learning models to learn mappings between surveillance cameras. In Chapter 4, a data-driven, bottom-up approach is used to exploit the wealth of information available on the Internet in order to achieve a representation compatible with the attribute representation introduced in Chapter 3. A fundamentally different approach is introduced in Chapter 5 which shows how to *transfer* previously learned models onto target camera-pairs using only partial annotation within the target domain.

- Chapter 6 formally identifies, and provides an initial investigation of, a novel direction for re-identification research using mobile re-identification platforms in lieu of static CCTV camera footage. The chapter re-examines common re-identification practices with consideration to a number of new challenges relating to the use of MRPs, and lays the foundation for an exciting new research direction.

- Chapter 7 concludes the thesis, discussing potential future research and extensions to the material presented in previous chapters.

Chapter 2

Literature Review

The standard scenario for re-identification is a finite network of surveillance cameras watching over public spaces through which people travel. Given a specific person of interest, nominated as a detection in one camera, the goal of re-identification is to locate that same person; ostensibly then, the aim is to retrieve by *identity* and not just appearance. Figure 2.1 on the next page illustrates this.

A basic pipeline for re-identification can be implemented using three major stages: (i) the acquisition of images of individuals from visual surveillance sensors (person detections), (ii) the generation of a representation of the person *e.g*. the person's signature (feature) and subsequent post-processes depending on experimental considerations such as memory and computational efficiency, (iii) and a matching stage, where a suitable method is applied between signatures to determine which are of the same person. Figure 2.2 on page 47 illustrates a more comprehensive re-identification system architecture, including extensions such as spatial sampling options and auxiliary information sources.

While person detection is the starting point for any re-identification system, for convenience the majority of re-identification work has assumed that perfect detections are readily available. Large bodies of research on person detection currently exist, therefore the reader is invited to examine Dollár *et al*.'s survey [44] for more detail in this area. Even with perfect detections, each of the remaining challenges presents significantly difficult questions to the re-identification community; what features are best? From which part of the detection should they be extracted and to which part should they be matched? Which matching strategy is best and under what

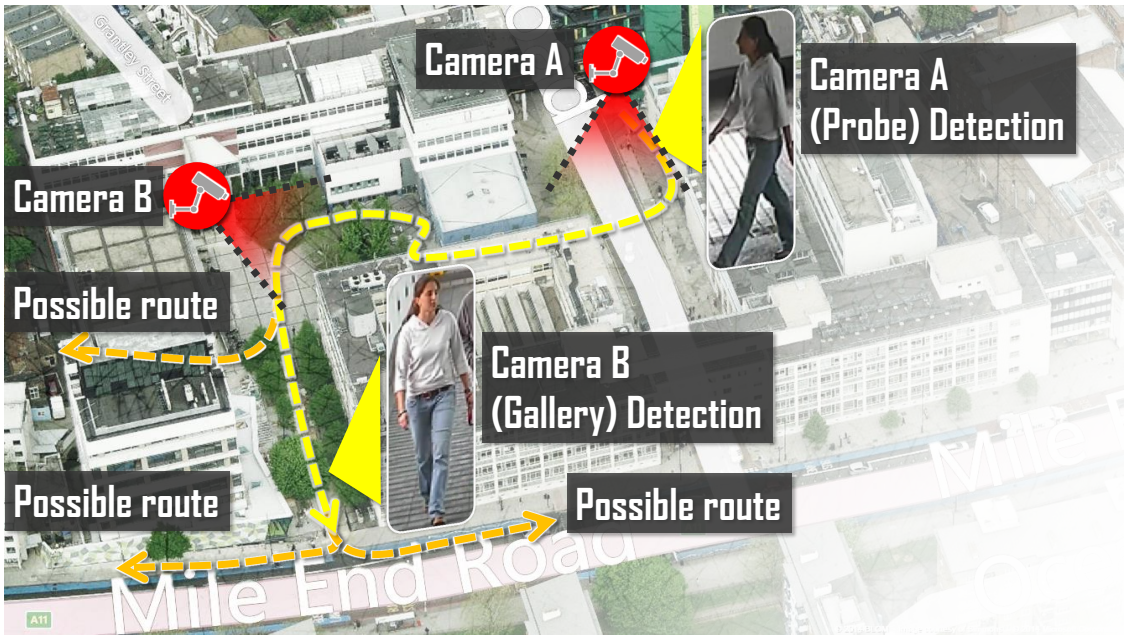


Figure 2.1: An illustrative example of re-identification in it's standard form; Two cameras (*A*,*B*) watch over some public space through which a person of interest travels and is selected as the *Probe detection*. The person then follows a route (yellow dashed), but when out of the camera view (red dashed), could take any conceivable route (orange dashed). Re-identification positively identifies the correct person as they appear in camera B.

---

conditions? This review enumerates informative work in re-identification that has contributed to answering these questions, and in the following sections we enumerate contemporary research relating to the approaches and results relevant to this thesis in later chapters.

Section 2.1 introduces re-identification research according to two main taxonomical axes that have emerged in the past decade, aiming to provide the reader with an insight into the challenges and responses from the re-identification research community. Section 2.2 examines a broad cross-section of attribute discovery and usage as useful background for Chapters 3 and 4. Section 2.3 examines work on transfer learning in order to give the reader some background context for Chapter 5.

2.1 Re-identification

Commonly in re-identification research the person signatures have been taken directly as low-level feature (LLF) descriptions, reflecting photometric properties such as colour [125, 33, 138, 142], geometric properties such as texture and spatial structure [122, 47, 147], or combinations thereof [47, 68, 61]. The principles behind using LLFs are those of simplicity and speed since

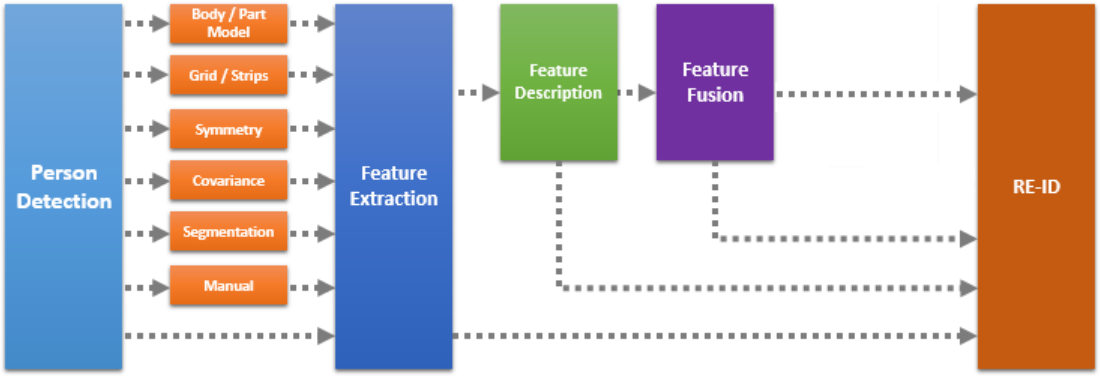


Figure 2.2: Feature extraction stages for a potential re-identification system; (i) person detection, (ii) diverse appearance decomposition choices permitting separate feature extraction from localisable body parts such heads, torsos and legs or uniformly sampled regions based on grid or strips, appearance symmetry-driven regions, covariance of the entire image, or segmentation or other manual selection; (iii) feature description, for example as a bag-of-words or histogram; (iv) fusion with other features; (v) dimensionality reduction and finally (vi) re-identification.

---

such signatures can be easily and reliably measured and also provide reasonable levels of inter-person discrimination together with inter-camera invariance. Once a suitable representation is obtained, nearest-neighbour [47] or model-based matching algorithms such as support vector ranking [147] can be employed to perform the matching and re-identification. In each case, the re-identification process is underpinned by a distance metric (*e.g.* Euclidean, $L1$-Norm or Bhattacharyya) chosen to measure the similarity between samples. Alternatively the distance metric may also be optimised [189, 77, 75, 76, 101, 103] or fused with auxiliary information [101, 103] in order to enhance the ability to find correct matches or to reduce mistaken matches, or imposters. A significant body of research has focussed on improving individual stages of re-identification [6, 101, 77], combinations of stages [47] and recently all-aspect re-identification pipelines [110].

Approaches to improving re-identification matching or representations may be categorised as (i) unsupervised (*i.e.* discovered from the underlying data structure) [61, 125, 71, 146, 40, 17, 15, 47, 32, 122, 123, 186] or (ii) supervised (discriminatively learned using labels) [68, 147, 179, 39, 188, 112, 128, 77, 41, 123, 107, 158], and lastly as (iii) engineered (procedural algorithms). However, it is also customary to describe re-identification systems with respect to the two common stages; namely improving either upstream (features) or downstream (matching task) performance. For example, both representations or matching algorithms may be engineered by "hand", or discriminatively learned. Typically, it is reasonable to expect better features to im-

prove downstream performance and thus overall re-identification performance. Synergistically, improving the downstream method can improve performance even when exploiting sub-optimal representations. In Figure 2.3 we illustrate this taxonomy using two orthogonal axes: *matching / representation*; and *engineered / unsupervised / discriminatively learned*, and discuss each in following sections.

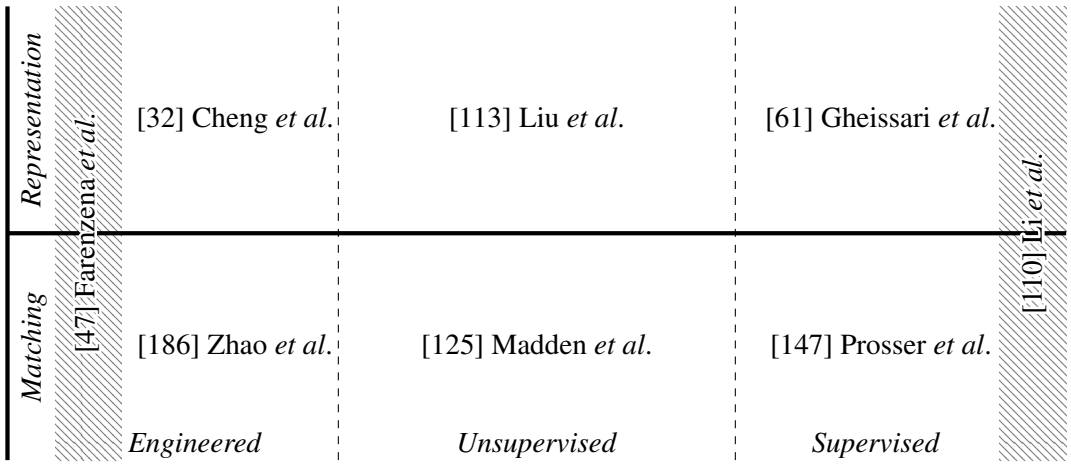


Figure 2.3: An illustration of two orthogonal contemporary avenues of re-identification research, novel representations and matching functions can both be "hand"-engineered, "learned" in an unsupervised sense via data-driven algorithms or discriminatively learned (supervised) using label data. Better representations embody intrinsically discriminative cues that are assistive for good re-identification performance; matching methods may be also be created or learned and improve downstream performance at the re-identification task. Some specific example classifications are given, where shaded regions indicate a work that belongs to both matching and representation categories.

2.1.1 Engineered Re-identification

The first sub-category of work embodies the view that practical challenges in re-identification can be approached from a purely practical perspective; the algorithm is directly engineered according to the insight and human expertise of the engineers. The majority of contributions to this category of research are low-level statistical features, but a significant proportion also attends to the open questions of how best to spatially sample the visual space and how to alleviate the issue of light-variations between cameras. Solutions range from sampling features as patches on a regular or overlapping grid, or as horizontal strips, through to exploiting second-order statistics for matching and body-part-model fitting.

Feature-design or feature-engineering approaches are a cornerstone of re-identification and often a popular first-generation research direction in other computer vision fields [164]. En-

gineered features may generalise more uniformly across camera-views as well as scale more tractably than discriminatively learned features trained with person detections labelled by identity [129] or pairs of detections labelled as being the same person [6, 7]. However, engineered representations for re-identification require no human supervision or annotation effort prior to use. It is extremely challenging to design unsupervised features that are both generalisable across all conceivable surveillance scenes, and robust toward practical covariates such as motion blue, clutter, view and pose change, dynamic lighting and occlusion; however this is precisely the goal in order to provide a high-performance feature. Another crucial factor is the question of which features and therefore which visual cues are "best". Colour is crucial in human visual perception, and is the most distinctive "low-level" feature [68, 70], however it is also one of the most prone to noise from the environment and may therefore be difficult to represent effectively for re-identification across an entire camera network. Most engineered features are *agnostic* to the matching method employed, thus may be matched downstream via nearest-neighbour distance metrics (*e.g*. Bhattacharyya, Mahalanobis, $L1$, $L2$, or cosine distance) prior to final re-identification; alternatively more complex matching algorithms may be used such as those discussed in Section 2.1.3.

The three forms of visual cue used to characterise human appearance for re-identification are colour, shape and texture. Although colour [145, 33, 138, 142, 6, 77] is an important cue, it is not discriminative enough to rely on alone and so other feature channels are often combined with them, such as texture and shape [68, 32, 47, 14, 186, 122] and more recently depth [3, 11].

Gray and Tao [68] exploit discriminative learning for feature selection in their early work (discussed in Section 2.1.1), however their engineered representation has been adopted by and has underpinned several works in the last decade. Their ensemble representation defined an all-aspect feature space comprising chromatic (RGB, HSV, YCbCr) visual cues as well as two families of filter bank responses (Schmidt, Gabor) for 19 additional texture channels. This ensemble of localised features (ELF, Figure 2.4 on the following page) space encodes a broad swathe of information. In Gray and Tao's work [68] the most informative channels are hue, saturation, blue, Schmidt filters, Gabor filters, and the red channel in roughly equal measure, however this uniform weighting of importance between features may not perform as well for a surveillance scenario where the global appearance trend is towards green clothing such as a military setting, or where blue and red apparel are worn only rarely.

Cheng *et al*. [32] also take biological inspiration toward re-identification, adopting Pictorial

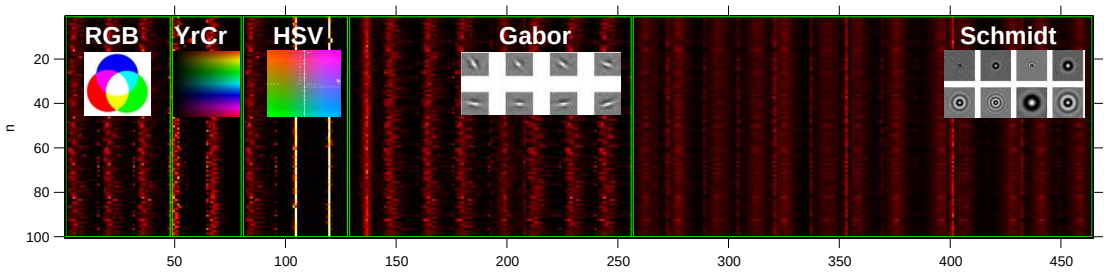


Figure 2.4: A depiction of 100 instances of ELF features. Feature channels for RGB, YrCr, HSV, Gabor and Schmidt filters are sectioned (green box) by their respective dimensionality. The full feature is 464 dimensions.

Structures (PS) for body pose estimation using a kinematic tree prior and local appearance representation from each inferred part part location. As in [47, 14], Cheng *et al.*'s method incorporates both chromatic histograms and stable colour regions computed using agglomerative clustering, granting a similar coverage of visual cues as Farenzena *et al.*'s work. Ma *et al.* [122] extract Gabor filters, weighted chromatic histograms and MSCR features from [56], applying covariance descriptors to the features extracted at several scales. The final representation is the concatenation of the differences between pairs of consecutive scales, unlike other covariance based approaches applied to probe and gallery combinations which cannot scale as efficiently. Although effective, the authors acknowledge no effort is made to learn more effective fusion weightings between their representation and the other features, and the approach assumes uniformity of texture in the background of each each person detection – an assumption that may be too brittle for detections of people wearing clothing that lack textured apparel or in street scenarios where the background indeed is more textured than the foreground.

Lastly, Zhao, Ouyang and Wang [186] propose a saliency-based method for re-identification that addresses both pose-variation and person-misalignment, a challenge in forthcoming real-world applications for re-identification that require automated person detection algorithms which may produce variably misaligned detections. Zhao's work looks toward mitigating this issue by detecting the most salient regions from detections in disjoint views and using saliency as a cue for matching between these salient regions. The features used include standard chromatic features as well as SIFT (Scale Invariant Feature Transform), which complements the colour histograms and is sampled densely over the detection in the form of patches.

The logical progression from matching single representations of human appearance is to partition appearance in some meaningful way in order to exploit spatial configuration and local

detail. Spatial decomposition strategies focus on either dividing the person detection into meaningful regions, such as with body-part localisation or body model fitting [21, 9, 47, 53, 52, 10, 186, 154] or spatial zones such as strips [147, 68, 87, 113], in order to facilitate inter-part matching [53, 47, 7, 101, 147, 142, 61, 31, 37, 10, 20] whilst reducing the effect of background noise on the representation, or to improve matching of misaligned person detections by selectively sampling only the most salient regions [186]. from the image based on some visual cue such as covariance [56, 122]. Patch-based methods sample on either an overlapping [77, 22, 23] or non-overlapping grid [106, 7], resulting in redundancy that may be exploited by dimensionality reduction in later stages.

Features that describe the entire image regardless of which pixel contains information about the person or background can lead to overwhelming quantities of background noise that can degrade final re-identification performance since typically matching is performed using the appearance information of the person, with background information a distractor. For example, even though the person depicted in two images may be in perfect alignment and be in the same pose and lighting condition, a change in background between the two images could alter the computed signature. Where automated person detection stages are employed, then human alignment cannot be guaranteed due to detection misalignment. More crucially, representations characterised as histograms discard important spatial information that may be useful for re-identification.

Various approaches attempt to circumnavigate these challenges and encode spatial information using a variety of strategies. Gray and Tao [68] and Prosser *et al*. [147] used an intuitive spatial model of horizontal strips, reasoning that re-identification data of the time consisted of arbitrarily positioned but horizontal viewpoints, thus vertically posed humans in horizontally aligned views would likely not benefit from the horizontal dimension. This confers the advantage of preserving a coarse spatial structure somewhat localised over the head, torso, legs and feet, but assumes that body proportion and alignment will be uniform between observations. Avraham *et al*. [6] employed a similar strategy with only five strips, and in later work used a discrete grid over both dimensions in order to capture potentially asymmetric appearance regions [7]. Tahir *et al*. [165] and Park *et al*. [142] manually define regions, such as upper or lower body, as does Huang *et al*. [81], noting that the strip-approach favoured in earlier literature is weak toward misalignment and therefore employing a heuristic aggregation of strips into a two-part model. Since humans often choose to wear one or two major articles of attire (such as a shirt

and trousers) these approaches indeed capture major areas of interest, however depending on the location and season, these cues alone may not be distinctive enough; for example as Gong *et al*. and others observe, during winter months populations tend toward uniformly dark clothing which results in heightened levels of ambiguity [104, 65, 132].

Recent work on face and scene alignment has enjoyed success in employing dense correspondence and likewise high-performance re-identification features have begun to exploit densely sampled patches from the source image. However, this strategy results in extremely high-dimensional features [77, 74, 23, 123, 156]. Most dense-sampling approaches employ a dimensionality reduction step such as principle component analysis (PCA) which selects the most variant regions; affording the latter a dense and rich selection of regions from which to select the most relevant for further processing [86, 15, 75, 77, 74, 178, 22, 110, 106]. Whilst these methods show promising raw performance at re-identification tasks, they are weak toward detection misalignment which can negatively impact re-identification system performance [186, 110, 65].

Person misalignment is a recently identified area of research in re-identification and to date there are only a handful of works that directly address it. Li *et al*. [110] implement an entire re-identification framework in a single deep learning network, exploiting the dropout trick [73] in order to force the network to randomly "forget" some patch displacement mappings learnt by the previous network iteration. This has the effect of preventing the network from overfitting – in this case, the network is prevented from forming debilitatingly strong mappings between feature filter banks determined by the previous layer, and their subsequent selection for, and mapping between, specific patches. In this way the final representation encodes a kind of soft spatial uncertainty for each spatial patch and most likely patch to test for a match. Whilst effective, this approach is unique to deep learning network design and therefore not available for most work. Zhao *et al*. [186] contribute a saliency-derived method for (i) identifying spatial regions of a person detection that are most distinct, or salient, and (ii) exploiting this cue by matching between detected salient regions. For example, given two detections of the same individual bearing a distinct green logo but otherwise uniformly dark clothing, the green logo will be detected as most salient so long as it remains visible, regardless of where in the detection's bounding box it is observed. A saliency-based method for re-identification that addresses both pose-variation and person-misalignment, an important forthcoming challenge for the next generation of real-world

applications for re-identification; these will require automated person detection algorithms which may produce variably misaligned detections. Zhao's work looks toward mitigating this issue by detecting the most salient regions from detections in disjoint views and using saliency as a cue for matching between these salient regions. The features used include standard chromatic features as well as SIFT (Scale Invariant Feature Transform), which complements the colour histograms and is sampled densely over the detection in the form of patches.

In addition to representation and sampling methods, the final family of methods pertains to the engineered matching of features between cameras. In an early example, Porikli [145] used correlation matrix analysis applied to whole-image colour histograms extracted from video frames to learn uni-modal colour transforms from a reference camera view to all other views in the network in order to "calibrate" the chromatic differences from one camera to the next; for example to compensate for light temperature differences commonly encountered when using mismatched camera equipment.

Gheissari *et al*. [61] and Madden *et al*. [125] both try to incorporate illumination invariance via normalisation strategies applied to dominant colours in order to build robust signatures. Gheissari *et al*. designed an algorithm specifically robust toward variable non-rigid clothing deformation over time. The method samples and segments multiple detections of the individual and then spatiotemporally cluster the regions in order to capture median chromatic appearance information from the major articles of clothing. Once this normalised, more stable colour information had been obtained, a manually defined decomposable triangular graph composed of vertices collected into triple-cliques is fit to the intermediary image. Since the arrangement of the graph regions is predetermined and ordered, the graph model permits trivial manual labelling as head, shoulder, torso, pelvis, thigh and lower-leg regions, although the authors employ an interest point detector and dense correspondence matching at the individual triangle/patch level. Although robust toward geometric clothing transformation over time, Gheissari's algorithm uses a simplistic colour representation that ignores potentially discriminative visual cues such as logos, detailing and trim. It also matches histograms extracted from a high number of triangular, non-overlapping regions which depend on initial lighting conditions, shading and does not generalise for all poses – for example where legs are in mid-stride.

Farenzena *et al*. [47, 14] address multiple challenges in their contribution, "Person re-identification by symmetry-driven accumulation of local features" (SDALF), which fuses

multiple basic features and direct matching methods into a single descriptor. Their approach to human pose variation is addressed with inspiration from gestalt theory [93], where a visual symmetry-driven body partitioning scheme underpins subsequent spatial sampling locations and provide cues for coarse pose estimation. The extracted features in this approach cover a broad range of information including a traditional histogram-based colour representation weighted by distance to the axis of symmetry for a coarse representation of visual appearance. Second, an aggregate colour descriptor similar to that employed by Madden *et al*. [125], encodes blob regions using agglomerative clustering of pixel data with respect to an inter-chromatic distance threshold in order to preserve somewhat more localised visual information. Third, Farenzena *et al*. construct a novel representation that facilitates matching and encodes recurrent patches present in the image, such as repetitive stripes, cheques or tartan. In order to achieve this, patches are randomly sampled from the detection close to the axes of symmetry, and thresholded based on the entropy of the chromatic content. Patches containing areas of low visual complexity are discarded since they likely represent areas of uniform colouration already encoded by the previous two features. Finally, patches are clustered by HSV colour and the centroid of each cluster is retained as the prime representation of the recurrent patch. Due to the different feature extraction strategies, distance calculation using SDALF is not uniform for each feature type, but are combined using a weighted sum at the final stage, making the SDALF compatible with other metrics and representations. SDALF is a compelling case, where a multi-faceted and multi-tiered approach to both spatial and visual cues results in a strongly discriminative representation without any need for arduous human annotation; however, it is also an illustrative example of a major flaw in this and other approaches that are modelled too narrowly and prone to dataset bias as it does not generalise to other common surveillance scenes uniformly [170].

Bialkowski *et al*. harness both the engineered discriminative abilities of SDALF and [172]'s regional covariance features in their re-identification work, applied to football. The work is a special case where half of the closed set of observable humans will be uniformly attired and intra-team re-identification is particularly challenging. Bialkowski's augmented each identity's visual model with non-visual cues as to the role of that particular player, exploiting a semantic non-visual cue for re-identification even though some players swapped roles throughout each match. In order to train the model, 25,000 labelled frames of field hockey data were used to discover role-order by formulating a linear assignment problem to be solved via optimisation.

Engineered representations and matching methods typically exploit either direct human-engineered, algorithmic approaches or second-order statistical methods such as covariance or information theory, a way of encoding direct human expertise as to what statistical information is relevant for the task; however in the next section unsupervised learning approaches are explored, which are related to engineered methods but include more complex activities such as clustering in order to discover latent structure within the data that can be exploited.

2.1.2 Unsupervised Re-identification

Like engineered representations, unsupervised representations require no human supervision or annotation effort prior to use, making this family of methods convenient and more scalable for the real-world needs of re-identification as a result. Engineered methods exploit human insight but one weakness shared by that family of methods is that they are *restricted* in composition to pre-defined concepts available to the engineer and do not take advantage of observable latent covariates within the data as practical cues. More concretely, they do not specifically address the presence of an underlying hidden structure present in the unlabelled data, that may be valuable to constructing a representation. Furthermore, in the engineered case the human engineer applies their intuition specifically with regard to realising some goal, for example re-identification. In the unsupervised case, there is usually no such assumption beyond the idea that uncovering a latent structure will be useful somehow.

Madden *et al*. [125] define a normalised distance metric in the RGB colour space where the Euclidean ($L2$) distance between two RGB triplets is normalised by colour magnitude. In the normalised space, colours are then manually discretised into "principle" colours, (interpretable as cluster centers in the aforementioned space), and enough principle clusters are retained to account for 90% of the pixels present in the image. To normalise the distribution of colour within each cluster, a k-means algorithm iteratively adjusts membership calculation and centroid adjustment and effectively "smoothing" the heuristically initialised cluster centers. In order to account for illumination variation between surveillance scenes (camera views), Madden *et al*. apply a colour intensity equalisation on both signatures; finally matching the normalised accumulative cluster distributions between people using Kolmogorov distance. Although robust toward geometric clothing transformation over time, Gheissari's algorithm uses a simplistic colour representation that ignores potentially discriminative visual cues such as logos, detailing and trim. It also matches histograms extracted from a high number of triangular, non-overlapping regions

which depend on initial lighting conditions, shading and does not generalise for all poses. The inverse is true of Madden *et al.*'s approach which identifies, refines and "soft"-matches major colour representations from the entire person detection.

Hamdoun *et al.*, and in separate work Khedher, Yacoubi and Dorizzi [91] like Gheissari, investigate interest-point detection-driven representation with a variations of the SURF interest-point detector [13]. In Hamdoun's work, the re-identification model is compiled over disjoint successive interest-points extracted from detections of people that are temporally spaced at even intervals and accumulated over time. The models are matched using sum of absolute differences (SAD) with efficient high-dimensional nearest-neighbour search made tractable using KD-trees [16]. Hamdoun's approach is unique in that it completely disregards colour, instead matching between the normalised distribution of first-order Haar-Wavelets found in the neighbour of each interest point. Again, the normalisation step confers a degree of illumination invariance, with the filter response aggregation encoding pose-invariant regions that describe visual commonalities between observations in different poses.

Liu *et al*. [113] exploit a data-driven approach to evaluating feature importance by learning a bottom-up measure and automatically adaptively weighting features according to the underlying appearance. Liu *et al*. address the question of which subset of available features should be used to best describe an individual observation a person, dependant on the background apparel and lighting for that specific observation. In order to achieve this, the authors cluster a given set of unlabelled images in order to discover *prototypes*, before assessing the feature importance within each prototype by performing graph partitioning. The application of a clustering forest-based method [26] for pairwise similarity estimation implicitly also provides feature selection and weightings that can be applied to new detections via assignment to one of the existing prototypes at test time.

In summary, a major advantage of data-driven approaches aside from not requiring explicit human labelling effort, is that models can be constructed from latent structure discernable from the data but not necessarily intuitive *a priori*. A slight disadvantage, is that a direct semantic interpretation of such latent covariates may not always be possible, or such an interpretation may not be as direct or immediate as that gained via discriminatively learned models (for example, as in the experiment results found in Chapter 4 as compared to those in Chapter 3). It is often desirable however, to have such a direct mapping between human expectation and algorithm

performance. The following section explores discriminative learning, or *supervised*, machine learning approaches to re-identification that seek to exploit human provided labels for this purpose.

2.1.3 Supervised Re-identification

Supervised methods have been applied to the task of modelling or "learning" representations, refining them via discriminative feature-selection, or inter-camera matching by way of learning a metric.

Learning an appropriate appearance representation for re-identification (also referred to in other re-identification literature as appearance modelling) is a popular area of re-identification research. Images can present intractable volumes of information and complexity at the pixel level, therefore it has become common practice to first concentrate solely on removing noise from visual information. Most appearance-based methods aim to extract relevant information in the form of global or local features. For re-identification, the goal is to provide inherently discriminative features that generalise well for unseen arbitrary surveillance scenes in real-world conditions and between different views, whilst inducing good performance during subsequent person identity matching.

In contrast to unsupervised methods, supervised learning approaches require manual human expert annotation cues to train a discriminative machine-learning algorithm to perform a task normally predicated on human experience and wisdom. Supervised methods are potentially more capable of being able to mitigate unconstrained misalignment and pose variations between detections observed in different views. However, doing so requires a trade-off between the cost and cardinality of human expert annotation on training data, the specific characteristics of the data, and the potential performance impacts on the trained model. Furthermore, such models may need additional training or retraining when applied to real scenarios or require additional annotation effort to compensate for more complex or different scenes, (*i.e.*, such models may not generalise from experimental conditions or specific camera views to applied conditions or other camera views).

In re-identification, supervised methods have been deployed for i) direct appearance modelling [154, 106, 20, 52, 182, 31, 110] or indirect appearance modelling methods such as appearance mapping methods [6, 146] and feature-relationship modelling [68, 52, 165]; ii) distance metric learning [39, 69, 179, 41, 128, 94, 77, 108, 27] or relativistic comparisons [147, 192]. For

representation learning in re-identification, the motivation is that stronger features are synergistic with, and thus capable of improving, discriminative learning model performance; therefore justifying the cost of expensive human annotation for the initial representation learning. However, such approaches may require additional processing or augmentation in order to generalise properly or be exploited for other camera views without additional annotation cost, retraining, or both.

Supervised representations in re-identification exploit human provided label information in order to learn functional mappings in the feature space that improve downstream performance by modelling characteristics inherent to a training set, the assumption being that the covariates learned by the model in this way will generalise to the test data. An early example of this can be seen in Gray and Tao's [68] work, introduced in Section 2.1.1; however since not every feature channel is equally contributive to the re-identification task the authors employ Adaboost to discriminatively learn a weighting on the feature channels. Adaboost accomplishes this by sequentially learning cheaply computable weak classifiers in a feed-forward multi-layered architecture. The learnt weights improve performance on the re-identification task, as well as being interpretable in some sense as a measure of each feature channel's overall utility for that task. In Gray and Tao's work [68] the most informative channels are hue, saturation, blue, Schmidt filters, Gabor filters, and the red channel in roughly equal measure. This determination is from the supervised learning framework employed to discriminatively learn the functional mapping that best permits use of the representation, however learning such a mapping may not be as useful for the case where a scenario exhibits a global appearance trend towards green clothing (such as in a military setting) where red may be an informative cue only rarely.

Prosser, Gong and Xiang *et al*. [146] build a bi-directional, cumulative brightness transfer function algorithm to robustly learn how chromatic cues map between disjoint camera views. By using a cumulative histogram for a descriptor, uncommon but useful cues from the underlying brightness distributions may be preserved, and contribute to the accuracy of the end mapping result. Colour is crucial in human visual perception, and the most distinctive "low-level" feature [68, 70]. It is also one of the most prone to noise from the environment and may therefore be difficult to represent effectively for re-identification across an entire camera network.

In other work, Prosser *et al*. [147] convert the re-identification task into a ranking problem where the goal is to ensure the correct matches between candidate pairs are ranked earlier than

incorrect matches. The motivation behind this change is ostensibly human in origin and inspired by expected real-world application use-cases. After the ranking process, a list of possible matches presented in order of likelihood can be quickly and more efficiently parsed by a human operator than if the operator must personally search through every possibility themselves through visual inspection. In effect, this is a feature selection process.

Li and Wang [107] also employ multiple intermediary learners, but instead jointly partition coarse intra-camera appearance (ostensibly, pose) and employ a semi-supervised approach to learn how to match between cameras. Since uncommon and more visually obvious appearances can be more easily matched by expert operators, the inter-partition scheme is effective for the more ambiguous and frequently encountered cases, where invariance is somewhat mitigated by matching like-for-like appearance configurations against each other and the possible transforms within each partition are less distinctive.

Mignon and Jurie [128] build a lower-dimensional space in which generality is preserved via sparse annotation, and into which person observations from two cameras may be jointly projected for more effective matching. Furthermore, their method applies the kernel trick to efficiently compute and map the data into a higher-dimensional, non-linear space.

Most recently, Li *et al*. [110] introduce a Filter-Pairing Neural Network (FPNN). FPNN is, uniquely, trained using several curated training strategies from deep learning to jointly learn a feature representation, invariance to geometric and photometric transforms between camera views and matching of identity using a novel dataset comprising of six disjoint camera views and 13,164 images of 1,360 people. The dataset, named after it's originating institution the Chinese University of Hong Kong (CUHK03), averages 4.8 images of the same person in each camera view. The network's convolution and max-pooling layers operate on colour patches sampled from paired images of human detections, thus applying pairs of convolution filter banks before representation as a vector of filter responses for each patch and contributing an analog to appearance transformation approaches mentioned previously. A separate patch matching layer matches patches between horizontal partitions from the images, learning patch displacement matrices that encode potential pose variations. This layer is subjected to a maxout-grouping step, with a winner-takes-all strategy that only updates the foremost activations from the precedent layer, effectively sparsifying filter responses and patch displacement. The process is repeated for a coarser grid of patches at a larger scale, affording the opportunity for the network to learn more discriminative cues at

different scales, similarly to recent work employing dense patch sampling and discussed in more detail in Section 2.1.1. Finally, the softmax function classifies the pair of detections. During training, the authors employ a curated range of strategies such as dropout, training image translation, data balancing and negative mining. Although effective on the CUHK03 dataset, validation on established datasets is not complete and FPNN was only evaluated on CUHK01 where it was outperformed by Kostinger *et al.*'s KISSME [94] at first-rank, and only performed comparably thereafter. Although the network learns a complex array of transformations and invariances, this does not guarantee performance on new, unseen surveillance views that may exhibit novel visual diversity. Furthermore, FPNN is computationally expensive to train and due to the complexity of the approach, may not scale to even larger volumes of data encompassing more camera views except in an "online" setting where the training is continuous, centralised, and long-term.

Supervised learning methods can also be used in lieu of making any assumption as to the nature of the distances between person signatures when matching; thus providing the means to learn a relevant and discriminative distance metric for the re-identification task. Distance metric learning (DML) can be used to infer either a global or local pairwise similarity metric from a set of labelled labelled images [179, 39], commonly setting equivalence constraints [77] and then formulating the task as a constrained convex programming problem [179, 41]. For re-identification, such constraints signify whether two signatures refer to the same entity (*i.e.* the same identity). Distinct from manually specifying a linear weighting for the combination of disjoint feature spaces [47], DML instead assumes a complex nonlinear space can be found to satisfy the re-identification task; aiming to maximise inter-person variation whilst minimising intra-person variation. Classic DML label information is normally coded as pairwise constraints on the data, being positive for equivalence constraints or negative for semantically dis-similar pairs.

Dikmen *et al*. [41] employ a Support Vector Machine (SVM)-style approach to obtain such a metric in their work using Weinberger and Saul's [179] large margin nearest neighbour (LMNN) classifier to learn a Mahalanobis metric that projects positive and negative pairs into a subspace where they are maximally distant. Dikmen *et al*. further extend [179]'s method by introducing a hard constraint for "imposters", that is, false matches falling within a certain distance from true matches are explcitly and forcefully rejected from the buffer zone by the cost function.

Hirzer *et al*. [77] apply dimensionality-reducing PCA step on a dense grid of sampled features

prior to metric learning in order to select the best spatial locations and reduce the complexity of learning a Mahalanobis metric using pairwise identity labels. However, a flaw in both approaches is that the Mahalanobis metric itself is linear, leading to suboptimal performance on particularly complex, non-linear data. Additionally, the computation of a full matrix that satisfies complex constraints and remaining a valid metric (*i.e*. positive semi-definite) can be intractable. Hirzer *et al*. reduce the dimensionality of the feature space and manually select the number of dimensions to retain using PCA, which permits a valid pseudo-metric to be computed. However the overall assumption is that the feature space can be reduced to a low dimensionality of $\tilde{3}0$, whilst retaining it's discriminative strength. Not all features are compatible with this assumption, particularly higher-dimensional features. Such features may be reduced to $\tilde{1}50$ dimensions but would require many more data instances than are normally available to construct a valid covariance matrix upon which to learn a valid metric. Although Hirzer's method is fast to execute for the standard benchmark datasets commonly used in re-identification research, it's reliance upon low-dimensionality mean it is not agnostic to all features.

Zheng *et al*. [188] reformulated the classic DML problem and minimised the probability that a true-match pair will be closer together than a false-match pair, in a similar relativistic approach to Prosser *et al*.'s RankSVM in [147]. Rather than operating directly to select or weight suitable features as is implicit in most DML and later ranking approaches, Zheng's method computes and exploits probability cues, a second-order property more robust and less computationally expensive.

Avraham *et al*. [6, 7] recast the standard re-identification problem as a binary classification task and uses a SVM trained on positive and negative examples of matching and non-matching human detections to train a discriminative re-identification classifier. Whilst effective at learning the appearance transition between pairs of cameras, Avraham's approach cannot scale practically to real-world systems since it requires a quadratic number of annotated pairs on the number of cameras in the network even assuming such transformations are commutative or bi-modal, and Prosser's [146], whilst bi-modal, uses whole-image representation in it's current incarnation, and does not actively select which modality to use when matching.

Supervised approaches to re-identification tend to perform better in comparison to either engineered or unsupervised algorithms in both representation and matching contexts, hence innovative work belonging to one of these families of approach are normally compared with other

members of the same category and not between categories. However, the cost of the generally superior performance gained by using discriminatively trained models and representations is the often substantial amount of human effort to provide sufficient volumes of data and labels with which train said models.

2.2 Attributes as Discriminative Cues

One view of visual attributes is that they are a form of semantic, transferable auxiliary or directly applied representation for higher-level vision tasks such as classification, recognition, description and retrieval [95, 97]. The use of attribute-based modelling for computer vision tasks is relatively recent, first proposed by Ferrari and Zisserman [51] and becoming widely employed in following years. One distinct property of most attribute research is that multiple attributes are employed in concert as part of an ensemble of standard machine learning detectors or classifiers that automatically assign human-semantic descriptive text labels to objects, entities or scenes.

Where attribute modelling differs from approaches that measure distances between low-level statistical features or discriminate between identities or large numbers of classes, is that attributes provide an intermediary basis that can assist in the high-level task by exploiting the low-level features in a different setting. In essence, attributes can provide an alternate vocabulary at an intermediary level, one that is inherently more expressive of higher-level semantics than the data used to train the attribute.

Typically in attribute learning therefore, it is key to address such questions as (i) what ontology of attributes to choose, (ii) how to learn them and perhaps most crucially, (ii) how to ensure the attributes are complimentary to each other for re-identification.

2.2.1 Ontologies and Attribute Discovery

For many years, ontologies were the subject of much debate and disagreement in the artificial intelligence community and in philosophy for much longer. One of the more elegant and concise definitions casts the ontology as a "formal, explicit description of concepts" [133]. But what concepts? Ontologies can be complex, multi-layered hierarchies or may be as simple as a flat list of classes, depending on the application area. Determining the ontology of an attribute-based system is often the foremost step, and a paradoxical one since we will not know which attributes will be informative for a given task. However, we can reason that human intuition and expert

knowledge will be useful, thus many attribute works tend to be motivated by human practices and wisdom. Another key issue requires mention here as well as in Section 2.1.3; following the definition of the ontology humans must annotate sufficient instances of data with which to train a classifier of some kind to a reasonable level of detection accuracy.

The majority of recent work on attributes looks to human expertise in answer to the question as to which attributes to learn [183, 173, 99, 51, 30, 100, 58, 141, 97, 96]. Where ontology selection is not performed manually it is discovered automatically from a data source [18]. Hand-picked ontologies can be thought of as being top-down and bottom-up. In the top-down case, ontology selection may be predicated on the knowledge of experienced human domain-experts. In the latter it may be based on the intuition of vision researchers, based on factors such as how detectable an attribute might be with available methods or data availability. There is a distinction between the selection of the ontology itself, and the discovery of an automated, data-driven ontology; expert-defined ontologies are subjective but may not include discriminative attributes beyond the experience of the expert, meanwhile data-driven ontologies are fundamentally not semantically based except insofar as they reflect some statistical property such as frequency [115].

Bottom-up attributes are often incrementally higher-order semantic terms as compared to the low-level representation they are learned from. Commonly, these kinds of attribute learning focus on geometry, texture and chromatic attributes such as *"red"*, *"striped"* or *"furry"* [51, 173, 100, 98, 184] or body-part localisation and classification of limbs, torsos, arms, heads and legs [119, 48]. Most works in these categories focus building a mapping from the semantic concept to visual pixel representations seen in the training data. In early work by Ferrari and Zisserman [51] the authors probabilistically model the properties of elementary attributes using their own intuition as to what humans find helpful in identifying and classifying objects, such as colour (red, green, blue, yellow) and pattern (stripes, dots, checkerboard)[1]. The configuration of neighbouring segments detections is used to infer slightly higher-order descriptions (stripes, spots). The resulting ontology of attributes is therefore one of intuition; the authors select standard primary colours as well as green as well as three of the most rudimentary patterns, but do not discuss their motives for doing so explicitly thus it is implied that these choices are based on

[1] In [51], attributes are categorised as binary and unary, and it may be helpful to the reader to contrast the authors terminology with contemporary usage to avoid confusion; authors use binary and unary to refer to the number of segments used to train attributes of each type, the contemporary definition of binary attributes refers to attributes that are either present, or not.

intuition. Another interesting observation about this work is that the authors heavily imply the possibility of attributes being a transferrable context, suggesting that learning stripes from zebras facilitates being able to learn stripy t-shirts. This theory is later directly examined by Lampert *et al*. [99, 100], who introduced an important dataset for attribute research using animals labelled according to expert findings from Osherson *et al*. [139].

Lampert, in a top-down approach, *et al*. annotated over 30,000 animal images according to the 50 classes and 85 visual attributes defined by Osherson *et al*. Osherson *et al*. introduced an ontology of animal "properties" (attributes), such as "nocturnal", "domesticated", "fierce", "eats plankton", of which many are visual in some sense, and employed undergraduate volunteers to review the list of animals and properties and annotate randomly chosen animals for the presence of the 85 properties.

Van de Weijer *et al*. [173] explore a different interpretation of transferable context without explicitly defining it as such, reporting that chromatic attributes learned from synthetic (though human selected) colour "chips" are less successful than those attributes trained using real-world photographs. However in some sense, the performance difference between synthetic and *verus-mundi* training data could be seen as an effect of the modality change between the stable and arbitrarily defined boundaries of the source colour spaces and the specific covariates present in the images used for testing, thus an artefact of the transfer problem explored further by some of the literature reported in 2.3.1. The authors select an ontology of colour attributes based on historic expert research by Berlin and Kay [19]; later, Kuo, Khamis and Shet also select the same ontology in [98]. Berlin and Kay present an ontology of colour attributes, partitioned discretely into linguistic terms that the authors posited was indicative of the culture's overall development. Various cultures were studied and observed to have evolved linguistic terms for their perception of colour and according to the evolutionary state of the dominant language, each culture shared commonalities in it's treatment of perceptual linguistic terms. For example, "stage 1" languages only distinguish between "dark/cool" and "light/warm" colours and later stages up to "stage 8" incorporating further colours in order, such as red at "stage 2", either green or yellow at "stage 3" toward distinguishing between black, white and grey, yellow and orange at "stage 7". Whilst interesting from an anthropological perspective, colour as a soft biometric has been employed to facilitate robust matching in the face of otherwise detriminental photometric variance in other work [98, 184]; neither of which can conclusively determine that colour attributes *alone* are dis-

criminative enough to be solely sufficient for good re-identification performance in real-world systems. The main concern is that multiple people can potentially share identical person signatures even when colour naming is applied to different items of apparel, thus more and potentially non-chromatic attributes must be obtained for this reason.

In Chen *et al*.'s work [30], the authors introduce a never-ending learner based heavily on previous work by Carlson *et al*., who define a process for iteratively refining an image description system with access to constant streams of new data. Carlson *et al*. initialise their system manually with an ontology of 123 categories of location (*e.g*. mountain, lake, city, museum), people (*e.g*. scientist, writer, politician), animals (*e.g*. reptile, bird, mammal), organisation (*e.g*. company, university), and miscellaneous others. Semantic relationships were also defined and part of the ontology, describing putative links between different categories; such as that books are written, and companies produce products. Each category was initialised with 10–15 seed instances and the system is left to run long-term, ostensibly because of current computational limitations. In both Carlson *et al*. and Chen *et al*.'s works, the goal is to begin with a small, seeded ontology of attributes amidst knowledge of classes and relationships, with a view to discovering more over time and the criterion with which Carlson *et al*. choose the initial ontology is again not discussed but assumed to be sufficiently broad to permit additional concepts and attributes to be discovered with every new iteration. Indeed, Carlson *et al*. report the first iteration results in almost 10,000 new "beliefs", and subsequent iterations resulting in fewer, around 1,000 new "beliefs". This suggests that many of the discovered beliefs, predicates and attributes are indeed shared between classes – particularly gratifying due to the scale of both works and in light of observations by Lampert *et al*. in [99, 100].

Attributes themselves are not necessarily a final representation, but may be further augmented by various strategies including other ontologies. Parikh and Grauman [141] cast attributes from being binary (present or not) to being relative to one another, for example being able to describe an image of a person as being "more smiley than ..." or "prettier than ..." another person. As in previously discussed work, Parikh and Grauman rely on ontologies from human experts and in this case the ontology is heavily inspired by or a subset of, ontologies released by Oliva and Torralba in [135] and Kumar *et al*. in [96]. Kumar *et al*. engineered their ontology based on the concept of "similes", using 65 visual attributes recognisable from face images, such as gender, race, age, hair colour and training such classifiers on vast quantities of human annotated data

obtained using Amazon's Mechanical Turk[2]. A key aspect to this research is the concept of similes, where images of specific regions (*e.g*. mouth, eyes) of individual "reference" people are first trained as weak learners in a simplified boosting framework as in [95]. The crucial idea is to define an auxiliary ontology/human semantic basis composed of facial region classifiers trained on images of specific people; this confers the ability to be able to classify how closely a new person's nose or eyes resemble the reference people and facilitates description by association such as describing a person's eyes as being similar to a particular celebrity (Figure 2.5 on the next page. The two distinct ontologies, relative and binary attributes, are complimentary in this work as well as remaining low-dimensional when compared to lower-level features.

Aside from top-down ontologies, bottom-up strategies can automatedly determine new attribute ontologies or intermediate representations using statistical methods to analyse data according to non-semantic assumptions. In a related work to Kumar *et al*., An *et al*. [4] use a reference set of people to augment such an intermediate representation. At *et al*. proposed that in addition to probe and gallery images, a separate reference set can be projected into a regularised canonical correlation analysis (RCCA) subspace that maximises the correlation between probe images from one camera, and gallery images from another. The reference set is thus likewise projected into the same space enabling a relativistic basis as an auxiliary data source to other, more low-level features and traditional re-identification matching techniques.

Van de Weijer *et al*. [173] explore a different interpretation of transferable context without explicitly defining it as such, reporting that chromatic attributes learned from synthetic (though human selected) colour "chips" are less successful than those attributes trained using real-world photographs.

To summarise, an ontology of visual classes may be generated by an expert (top-down definition), or be discovered after mining a sufficient quantity of data. In the former case, it is often not possible to determine the effectiveness of a given ontology *a priori*, a disadvantage since there is no guarantee sufficient positive instances of a given attribute may be present in the available data with which to train a capable visual classifier. In the latter case, this problem can be avoided with a bottom-up, data-driven approach to attribute selection but at a penalty to how *semantically interpretable* the resulting attributes are to humans. The following section examines the automatic discovery of attributes, which can be viewed as a bottom-up way of determining a type of

[2]https://www.mturk.com/mturk/welcome

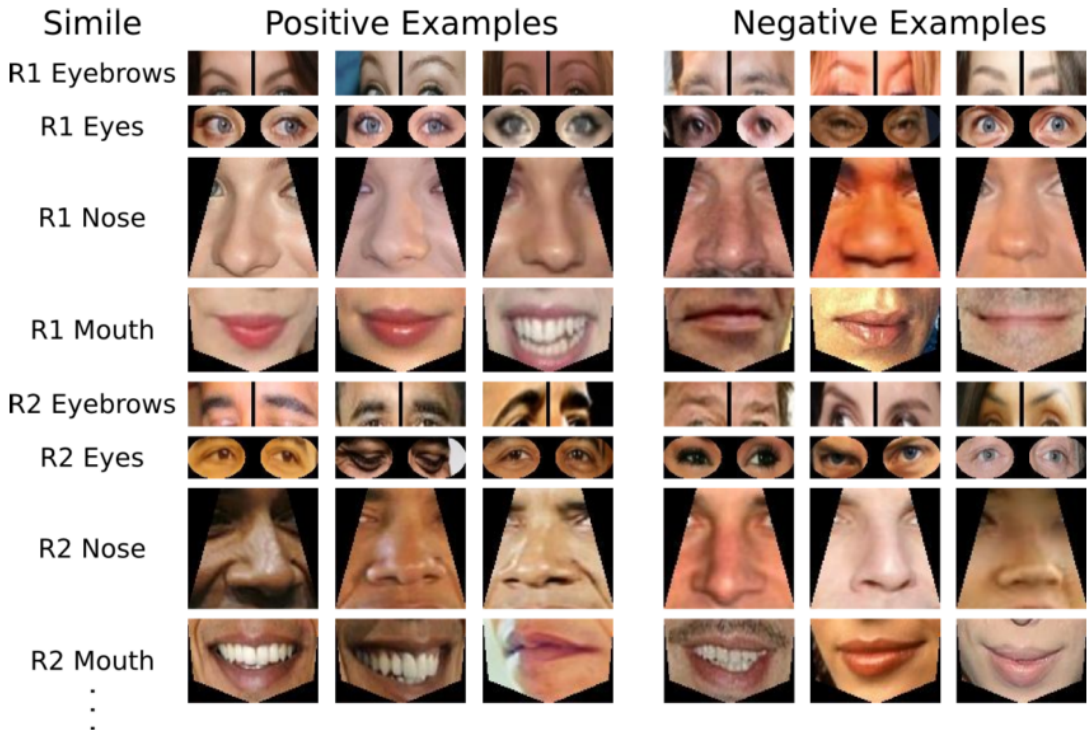


Figure 2.5: Examples of "simile" classifiers from Kumar *et al.*'s work in [96]; Similes are a form of auxiliary ontology capable of augmenting more traditional and direct attributes such as "hair colour" or "gender".

ontology as well.

2.2.2 Discovering Attributes Automatically

Attributes have been discussed in the context of being learned within a supervised framework, where the semantic knowledge has been provided by human expertise. However, human expertise is finite in many respects and this approach requires an ontology of attributes as well as sufficiently labelled instances to train classifiers for them. Sufficient quantities of labels are not always available or available in sufficient volume (*i.e.* may be sparsely represented in the data), or may contain annotation bias or errors. Furthermore, the human defined ontology will always be intrinsically incomplete in the sense that it cannot be guaranteed to provide enough discriminative information by itself to complete the task perfectly for every imaginable instance. Most crucially it is impossible to determine how well the ontology can be classified *a priori* to training.

A collection of latent attributes may also be discovered in an unsupervised sense, by mining data. In this case the domain-specific basis-set to be discovered normally conforms to statistical properties that maximise the performance of the system on some task. An example of this form of

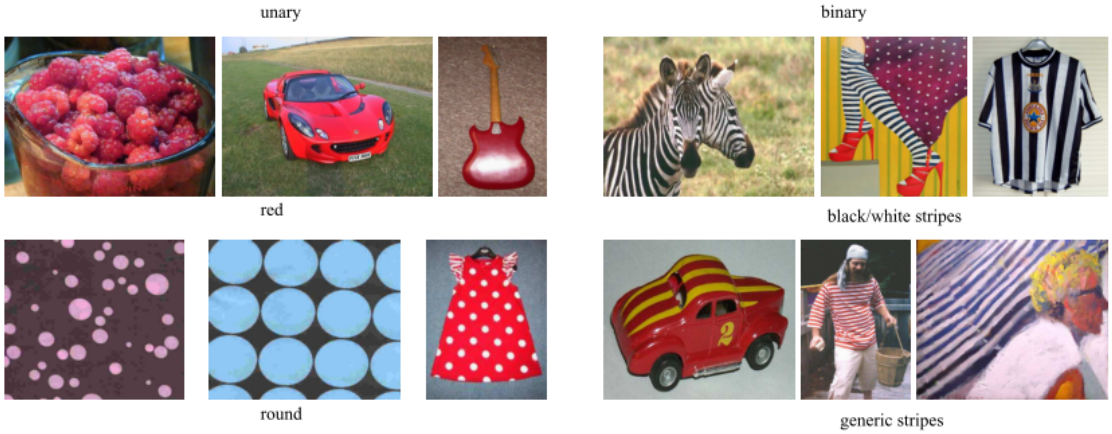


Figure 2.6: Example training data used to construct classifiers capable of recognising elementary visual attributes in Ferrari and Zisserman's work [51]; images are decomposed into segments, whereupon unary attributes like *red* may be learnt from entire segments and more complex attributes such as *stripes*, (right) may be learnt from two (binary) neighbouring segments.

attribute discovery can be seen in principle component analysis (PCA, [85]), where the basis set is selected such as that the maximally variant dimensions are the most useful bases upon which to project the data. Several unsupervised bases have been successfully employed for dimensionality reduction and thus could be considered domain-specific attributes, where the bases favour some property such as variance (*i.e.* principle component analyses, PCA [85]), or frequency (*i.e.* topic models [24]), sparsity [55]. Neural networks can be taken as a further example. Supervised approaches such as multi-layer neural network modelling [80, 110] learn to approximate continuous functions in vector spaces using combinations of locally derived bases. Essentially, expert defined annotations are used to train the network and the hidden layer of the network eventually becomes a complex set of attributes that facilitates the network's performance on some task. However, human interpretation of neural network layers is usually subjective, since no guarantee is possible that the learned weights will have any definite semantic content or meaning.

Efforts to automatically learn semantically-meaningful ontologies and attributes exist and normally seek to exploit existing bodies of information such as curated websites (*e.g.* for domain-specific knowledge) or even the open Internet. Berg, Berg and Shih [18] take this approach in order to discover visual attributes from retail product imagery depicting items of luxury apparel; the aim being to determine (i) which semantic text in the product description describes (ii) which region of the item depicted in the image (Figure 2.7 on the facing page).

So far, research encompassing both data-driven and human expert-defined attributes is rela-

Figure 2.7: Examples from Berg, Berg and Shih's [18] automatically discovered "handbag attributes", ordered by *visualness*.

tively sparse in comparison to solely data-driven or solely human expert-defined attribute discovery methods. Liu and Kuipers [115] develop a unified framework for action attribute recognition, where the ontologies of attributes employed are either manually defined or learned from the underlying data. Since the two types of attribute are disparate, the potential distribution differences between them could potentially be severe, however Liu and Kuipers employ a Latent SVM framework similar to that used in [50] which simultaneously learns attribute and weighting together thus addressing the problem illustrated in the previous section.

In [59], the authors also take the same approach and augment an ontology of human expert-defined attributes with support from data-driven attributes. The authors learn a unified *semi-latent attribute space* which represents the joint space of human-defined attributes as well as capturing the natural structure and properties of the data that are not already included in the human expert's attribute definitions. Fu *et al*. further extend their ontology with a third, class-conditional attribute type, inspired by [78].

To summarise, an ontology of attributes may be generated by an expert (top-down definition), or be discovered after mining a sufficient quantity of data. In the former case, it is often not

possible to determine the effectiveness of a given ontology *a priori*, a disadvantage since there is no guarantee sufficient positive instances of a given attribute may be present in the available data with which to train a capable visual classifier. In the latter case, this problem can be avoided with a bottom-up, data-driven approach to attribute selection but at a penalty to how *semantically interpretable* the resulting attributes are to humans.

2.2.3 Attribute Informativeness and Reliability

Although attribute reliability can be quantified by measuring the reported error against known test data, *detectability* and *discriminativeness* (*i.e.* informativeness) are factors that may present significant challenges for any "downstream" machine learning tasks reliant on attributes. Inter-attribute interference may affect overall system performance on many predictive machine learning tasks to degrees that are *a priori* impossible to predict meaningfully before training time. One admittedly perfunctory analogy available to us when describing a system that utilises intermediary representations such as attributes as part of a multi-layer model, is that of the neural network. In the purely data-driven attribute case, the analogy is stronger still, since the data-driven attributes are not necessarily semantically useful but provide a useful basis for the accomplishment of some task - however where neural networks exploit algorithms like back-propagation in order to learn the correct weight assignments for each "attribute" in the hidden layer, many modern methods using attributes in this setting do not [99, 100] and thus the error inherent to the "raw" attributes propagates downstream and penalises task performance [148, 166].

There can be no assurance of a particular ontology affording the level of discriminativeness required for a theoretically "perfect" re-identification or retrieval system. As well as this, attributes taken alone do not offer enough of a cue for automated re-identification or retrieval tasks, hence they are used in collections; however when using multiple attributes in a collection, score calibration becomes a significant concern. Simple concatenating or the stacking of attributes together as in [100] makes a significant assumption; that the raw attribute scores are all uniformly distributed; which in fact may be far from true; additionally, methods that subsequently rely on matching by attribute vector similarity are then ill-posed since distances will not conform to any meaningful space, but rather to multiple overlapping and localised subspaces. Other work operating on the assumption that attribute scores are Gaussian distributed include Siddiquie *et al*. [159] and Zhu *et al*. [193]. Zhu *et al*. focus their efforts on reducing annotation ambiguity in their manually selected ontology of attributes, only accepting a ground-truth annotation should it be

supported by three votes and otherwise only accepting the assignment as tentative; however this work solely investigates attribute classification performance using boosting methods and does not directly make any effort toward ensuring the attributes are calibrated for further use. In Siddiquie *et al*., the authors investigate retrieval of faces via attributes using a ranking framework that infers a set of additional discriminative attributes that support the initial query and incorporates learning the relative weightings of the support vectors as part of the model. Finally, the previously mentioned work by Liu *et al*. [115], utilises a latent-SVM framework that likewise jointly learns the correct weightings required to mitigate any downstream issues arising from the uncalibrated "raw" attributes simultaneously with other variables.

One contrarily intuitive example of detectability, is that visual attributes such as "blue" or "stripy" are not *de facto* more accurately predictable than non-visual attributes *per se* [100], and non-visual attributes can themselves be learned to a high degree of accuracy in some cases where enough visual correlations exist to support subsequent learning [100]. To further illustrate this, a visual classifier may be able to predict an animal as being "smelly" if it resembles the visual characteristics of a skunk, despite there being no direct visual cues alluding to the property itself.

2.3 Transfer Learning

A central limitation of most existing discriminative learning approaches is that they are only tractable on closed-world benchmark problems than realistic volumes of data and real-scale scenarios. In particular discriminative learning methods often require many labelled instances for training, a potentially costly and time-consuming process. An additional assumption is that this training data must be from a *target* data domain, in order to learn and exploit the practical covariates and distributions as present in the application domain and which are unique to that application domain. In essence, the transfer learning task is to mitigate the distribution disparity between domains. This is reasonable for training or testing splits on benchmark datasets that are already exhaustively annotated by person identity or potentially for static re-identification systems that will never be moved and consisting of few cameras. However it is highly impractical for real-world use, where there may be very many pairs of cameras in a given network, *each* requiring exhaustive annotation – and new cameras added over time. Therefore, such prerequisites would render scaling systems to useful real-world levels would be impossible or prohibitively expensive. Ideally, we would like to deploy a re-identification system between a pair of cameras with

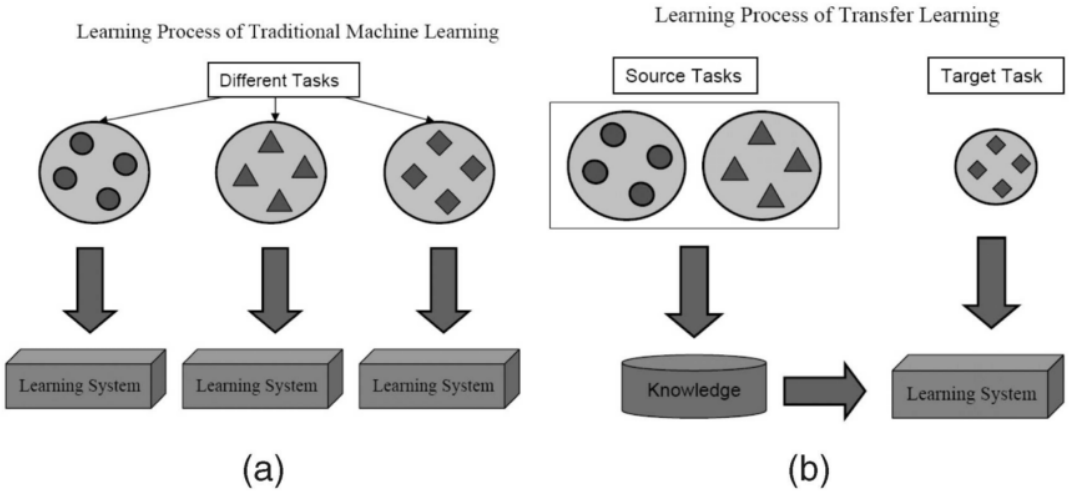


Figure 2.8: Learning processes for (a) common machine learning pipelines and to contrast, (b) transfer learning pipelines, as presented in Pan and Yang's survey of transfer learning research in [140]

---

minimal calibration/training annotation. What a system learns from annotations of one camera pair should be exploited by another pair without requiring exhaustive annotation in the new pair.

This is a main motivation for *transfer learning* [140, 45, 83], which is already important for many classical vision problems such as object recognition [151] with multiple classes or domains. However it is critically important for re-identification because the number of domains (camera pairs) is *quadratic in the number of cameras*, thus Section 2.3.1 discusses this specific application area in further detail. Obtaining exhaustive training data for every domain is even more impractical than for conventional vision applications, hence transfer learning becomes a critical tool in avoiding this obstruction. In the following section, relevant background work is presented to the reader as context for Chapter 3 which benefits from an initial commentary on transfer learning and recent works within this field; secondly, in Section 2.3.1 we enumerate a narrower selection of works exploiting transfer learning to good effect specifically for re-identification.

Transfer learning [140, 45, 83] has been used in numerous classical computer vision problems, for example object categorization [83, 151]. The motivation is typically to scale systems to many classes [83] or domains [151, 45] without requiring prohibitive amounts of additional training data for each new class or domain. While transfer learning is already an important issue in classical vision tasks, it will turn out to be even more central to the re-identification problem. This is so since *pairs* define domains in this context, thus it is highly impractical to collect

exhaustive training data for a quadratic number of domains.

Transfer learning research can be neatly categorised as belonging to one or more of three distinct research areas; determining what to transfer, how to transfer and when to transfer [140], of which the most relevant to this thesis are the initial two in particular. Pan and Yang [140] further summarise work on determining what to transfer as: instance-transfer, feature-representation-transfer, parameter-transfer, or relational-knowledge-transfer problems.

One of the most simple transfer learning techniques but perhaps one of the more sensitive to the sheer volumes of data available are the instance transfer methods [36, 111]. This setting assumes that sufficient data co-exists in both classification domains, that the intersection, or co-current data, may be reused for the target domain by learning a new weighting for the source domain. Dai *et al*. [36] do precisely this, by sparsely annotating data from the target domain and identifying training instances that co-occur and with what distribution. By re-weighting training data that has been discovered to violate the traditional AdaBoost assumption of identical distribution among training and test data, the impact of data from the transgressing distribution can be controlled and the effect of negative transfer mitigated. Lim *et al*.'s [111] contribution likewise "borrows" specific instances from visually similar classes but takes the further step of applying transformations to them in order to synthetically alter the training examples to be more informative about the target class, for example stretching images of *armchairs* to better visually resemble *sofas*, or learning that "toilets" resemble aspects of "cups" and "saucers". However, the assumption this approach makes is that a sufficiently varied collection of object classes is available and labelled, such that there are sufficient data sources to reliably generate new synthesised instances for the target class is definitely available for transformation, and that such transformations successfully resemble new instances of the target class. One may envisage this approach working well for say, cars or other rigid artefacts, but less well for deformable objects, or for classes with much more extreme intra-class visual variations.

One of the classical motivations for transfer learning is the avoidance of onerous annotation work although most transfer methods still require at least some annotated data to work with. One of the first works in this field was contributed by Li, Fergus and Perona [49], where transfer learning was achieved by constructing a Bayesian prior from a generic model learned from set of objects. The previously encountered, *known* model forms a kind of intermediate representation which was exploited to describe a novel target model's parameter distribution. Li *et al*. achieve

this with a simplistic and generalised prior determined from just three initial classes, thus their approach may only improve on target classes with some relevance to the prior. Another open question with this work is whether other priors may be more useful, and if so, how to construct them and from what data? Another form of model-transfer was suggested by Tommasi *et al*. [169], that sought to address these questions some years later. Tommasi *et al*. considered multiple separate source models and introduced a discriminative approach to determine the linear combination of source classifiers that best describe the novel target class of data, thus addressing the principle questions of what to transfer and simultaneously calibrating (*i.e*. determine relevance weighting) how much to transfer from each source. Tommasi *et al*. further tune the transfer process according to specific performance at the target task, however the authors do not examine the same approach for tasks other than recognition, such as retrieval.

Feature-centric transfer methods construct representations that are robust to *inter-domain* variation, whilst preserving statistical or geometric properties useful for computer vision tasks, or for which inter-domain mappings can be computed in order to minimise differences in either marginal or conditional distributions. The key task is normally to discover some combination of previously computed features or property of features that assists in the target domain. Ruckert and Kramer [150] do so by treating this task as a *meta-learning* problem. The authors first proceed by determining kernels and their parameters for each source domain as with a standard kernel classifier but note that this standard approach leads to full rank kernel matrices due to aggressive regularisation and therefore the balance between generalising to new domains and remaining discriminative on the source domain is very much biased toward the latter. To avoid this the authors adjust the standard SVM learning paradigm and kernelise a form of cross-validation that ensures a restricted pool of potential kernels for the optimisation step, but provides an evaluation step with the entire source domain so as to encourage generalisable solutions. The differences between the learnt kernels can then be exploited to generate a kernel and classifier for the target domain.

Typically, most transfer methods including Ruckert and Kramer's require labelled target data. However, Long *et al*. [118] posit a method for representation learning that accomplishes these goals without the requirement for labels on the target domain. In order to achieve this, the authors adapt the joint distributions such that the expectations are matched between domains, however this is a nontrivial goal where there is no labeled data available. The authors reduce the

difference between both conditional and marginal distributions, since minimising the difference between conditional distributions does not explicitly do the same for the marginal distributions. Because of the lack of labels and therefore no discriminative statistics available on the target, the authors approximate using "pseudo labels", or labels obtained by blindly applying source domain-trained classifiers on the target domain. In essence the assumption is that the target domain's class-conditional distribution of *pseudo-labels* and source domain's class-conditional distribution of ground truth labels, as well as the source and target marginal distributions can be pulled closer together whilst ensuring the target domain's variance is maximised, thus ensuring the preservation of the properties that assist in classification tasks.

2.3.1 Transfer for Re-identification

In this section it is helpful to clarify that we consider a *pair of person detections* to make up a *domain*, and this should not be confused with some other studies which consider a particular *camera* to be a domain [151]. This consideration implicitly represents the fact that we require a visual appearance mapping function between detections obtained from distinct camera views. For classification [151] and detection [45], an individual camera encompasses the notion of a domain because a camera's parameters impart a systematic impact on the observations, which the model must learn to interpret. However in re-identification, the task for transfer learning is to infer something about pairs of observations, and the systematic impact of each dynamic scene on person appearance is therefore defined by the pair of cameras.

The pertinent issue in transfer learning [140] is the question of where to transfer *from*. When there is only one source of information available, and that source is known to be highly relevant to the task of interest, then transfer learning is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying and potential relevance. In this latter case, it is non-trivial to design models which avoid negative transfer [140]. Our problem of transferring mappings across camera pairs falls squarely into the latter more difficult case. Since the relevance of one camera pair to another depends on similarity in their viewing angles and lighting, many pairs will not be similar and working out which source is best to transfer from is of critical importance.

Only very recently has transfer learning for re-identification begun to be considered [109, 191, 105, 121]. However these studies mostly consider only improving within-domain (camera pair) re-identification by transferring knowledge learned from one group of people to help iden-

tify another group of people. This is intrinsically a much more restricted scenario than the more general and useful case of transferring across domains to permit re-identification in a new camera pair with sparse annotations.

Wu *et al*. [105] present the first general investigation of transfer learning for re-identification on a range of datasets. The authors argue that two pre-conditions must be met in order for feature-transfer approaches to help re-identification; source and target domain tasks must be related and sufficient observations of each "class" must exist, from which we infer the authors mean that multi-shot re-identification may be the main beneficiary of this family of transfer methods. Secondly, instance-selection methods are more beneficial than directly trying to learn an appearance mapping function, and thirdly that insufficient data in the different domains necessitates methods that thrive on sparse data or reduced numbers of annotations available.

Ma *et al*. [121] likewise investigate transfer learning for re-identification, employing a strategy reminiscent of aspects of Long *et al*.'s work [118]. The motivation behind this is to avoid extensive annotation of positive and negative pairs of individual person detections by modelling only the negatives which can be easily generated (and which are far more numerous), and estimating the positive-pair model parameters rather than learning them discriminatively then exploiting the assumption that the difference between positive and negative models will be similar for both domains. Interestingly, the work shows that the estimation error is invariant to the true means of positive and negative pairwise data from the source and target domains respectively; where the error can be bounded by negative instances only. This avoids the requirement for exhaustive annotation and enables a source model to be transferred to the target domain without onerous annotation cost.

Li *et al*. [109] highlight the fact that many works pursuing transfer learning for re-identification-related applications assume that the target domain consists of a sufficient quantity of data so as to be representative of the ongoing operation of the camera from which it is drawn; this is particularly salient since it exposes one assumption made by transfer-learning approaches for re-identification, *i.e*. that for real-world use these techniques require "chunks" of data extracted over time, and are thus applied to similar chunks which precludes their immediate use in real-time.

In Zheng *et al*.'s [191] work, the authors redress the problem as a binary verification task rather than the traditional person re-identification problem, where verification refers to whether the query person is on the watchlist or is just one of many unknown candidates (imposters). This

approach makes a strong assumption; that the re-identification task is specifically the watchlist verification task. Furthermore it is limited twofold; (i) it operates on individual query probes one-at-a-time, rather than being able to operate on all possible query probes in a "batch", and (ii) requires discriminative training on each pair of cameras. This work differs to most current transfer learning research in that rather than explicitly transferring knowledge of appearance, it aims to transfer a bipartite ranking function based on the difference between target and non-target person detections in each domain. The goal, therefore, is to directly predict a ranked candidate set containing the targets by exploiting second-order statistics (mutual information) rather than directly operating on appearance.

2.4 Summary

The dramatic rise in surveillance data volume has created a substantial deficit that has yet to be fully addressed but for which the application of machine learning, computer vision-based algorithms is almost certainly the only tenable solution. Recent research has enjoyed success but on closed-world, densely annotated scenarios [47, 186, 94, 6] where discriminative learning methods can leverage human expertise and with sufficiently diverse data, model requisite higher-level concepts such as binary attributes with some success. However, with consideration to the wide variety of practical covariates inherent to surveillance scenes, as well as the degree of challenge presented by the full range of all possible human appearance variations, hand-crafted low-level features cannot maintain sufficient performance across all possible real-world surveillance scenes and there is no intuition for the human operator who must parse the results, presented directly from the feature-space. Furthermore, representations for re-identification are often engineered with specific scenes or data in mind and do not generalise, or with specific assumptions as to the importance of each visual property that likewise do not hold for each new surveillance scene. Discriminative methods can be trained to yield substantially more suitable visual representations that lend themselves better to the downstream task of re-identification, however this relies on the availability of suitable volumes of both data and annotations to be successful. Discriminative methods can also be used to explicitly model the appearance change between cameras, however they are rarely scalable in terms of being able to deploy such methods onto large CCTV camera networks such as those seen in the real-world. The dominating reason for this is the quadratic amount of human annotation effort as the number of cameras rises, particularly apparent for those

methods that rely on pairwise training data in order to explicitly model inter-camera appearance change [6, 94]. Finally, the re-identification community has focussed on a standard formulation of the re-identification task that assumes at least two cameras and a closed set of probe images and gallery images, with the expectation of a one-to-one match being possible for all probes. In the real world, these assumptions may prove too strong due to an arbitrary number of "imposters" being present in the relevant gallery set, the availability of new sources of surveillance data featuring different visual covariate properties and dynamic environments where modelling every ingress and egress point is impossible.

This thesis addresses such challenges. It introduces a new human-semantic, visual representation in Chapter 3 that bridges the semantic gap between human operators and surveillance task operational requirements. In Chapters 4 and 5 two distinct methods for addressing the challenge of scaling surveillance systems to fulfil real-world requirements are developed and reported. In Chapter 6, the question of whether such systems could function for mobile re-identification platforms is raised and we break new ground by introducing a novel dataset and preliminary study of this exciting new research direction.

Chapter 3

Human Attributes

In this chapter, we take inspiration from the operating procedures of human experts [132, 174, 32] and recent research in attribute learning for classification [99] in order to introduce a new mid-level *semantic attribute* representation of humans that incorporates view invariance between public spaces, "zero-shot" queries for cases where an initial visual observation is unavailable, and moreover we show how to complement this new representation with a simple discriminatively trained metric.

In order to initially define what we refer to as attributes, we observe that when performing person re-identification, human experts rely upon matching appearance or functional attributes that are discrete and unambiguous in interpretation, such as hair-style, shoe-type or clothing-style [132]. This is in contrast to the continuous and more ambiguous quantities measured by contemporary computer vision based re-identification approaches using visual features such as colour and texture [67, 147, 47]. This attribute-centric representation is similar to a description provided verbally to a human operator, e.g. by an eye-witness. We call this task attribute-profile identification, or *zero-shot re-identification*. Furthermore, we will show in our study that humans and computers have important differences in attribute-centric re-identification. In particular descriptive attributes that are favoured by humans may not be the most *useful* or *computable* for fully automated re-identification because of variance in the ability of computer vision techniques to detect each attribute and variability in how discriminative each attribute is across the entire population.

This approach of measuring similarity between attributes rather than within the feature-space

has two advantages: (i) it allows visual re-identification (from a probe image) and semantic identification (from a verbal description) to be performed in the same representational space; and (ii) as attributes provide a very different type of information to low-level features, which can be considered a separate modality, they can be fused together with low-level features to provide more accurate and robust re-identification.

3.1 Problem Definition

3.1.1 Attributes as Representation

Attribute based modelling has recently been exploited to good effect in object [99] and action [115, 58] recognition. To put this in context: in contrast to low-level features or high-level classes or identities, attributes provide the mid-level *description* of both classes and instances. There are various unsupervised (e.g. PCA or topic-models) or supervised (e.g. neural networks) modelling approaches which produce data-driven mid-level representations. These techniques aim to project the data onto a basis set defined by the assumptions of the particular model (e.g. maximisation of variance, likelihood, or sparsity). In contrast, attribute learning focuses on representing data instances by projecting them onto a basis set defined by domain-specific axes which are semantically meaningful to humans. Recent work in this area has also examined the exploitation of the constantly growing semantic Web in order to automatically retrieve visual data correlating to relevant metatext [51] and vice-versa for visual retrieval using metatext queries [153].

Semantic attribute representations have various benefits: (i) in re-identification, a single pair of images may be available for each target – which can be seen as a challenging case of "one-shot" learning. In this case attributes can be more powerful than low-level features [99, 159, 115] because they provide a form of transfer learning as attributes are learned from a larger dataset *a priori*; (ii) they can be used synergistically in conjunction with raw data for greater effectiveness [115]; and (iii) they are a suitable representation for direct human interaction, therefore allowing searches to be specified, initialised or constrained using human-labelled attribute-profiles [99, 159, 97].

3.1.2 Attributes for Identification

One view of attributes is as a type of transferable context [189] in that they provide auxiliary information about an instance to aid in (re-)identification. Here they are related to the study of

soft-biometrics, which aims to enhance biometric identification performance with ancillary information [82, 38]. High-level features such as ethnicity, gender, age or indeed identity itself would be the most useful to us for re-identification. However, soft biometrics are exceptionally difficult to reliably compute in typical surveillance video as visual information is often impoverished and individuals are often at "stand-off distances" as well as in unconstrained or unknown viewing angles.

Alternatively attributes can be used for semantic attribute-profile identification (c.f. zero-shot learning [99]), in which early research has aimed to retrieve people matching a verbal attribute description from a camera network [174]. However, this has only been illustrated on relatively simple data with a small set of similarly-reliable facial attributes. We will illustrate in this study that one of the central issues for exploiting attributes for general automated (re)-identification is dealing with their unequal and variable informativeness and reliability of measurement from raw imagery data.

In this chapter, we move towards leveraging semantic mid-level attributes for automated person identification and re-identification. Specifically, we make four contributions as follows. (i) In Section 3.2.1 we introduce an ontology of attributes (see Table 3.1 on page 83) based on a subset from a human expert defined larger set [132]. These were selected for being relatively more reliable to compute whilst also discriminative for identification in typical populations. (ii) We evaluate our ontology from the perspective of both human-centric and automation-centric purposes and discuss considerations for successful ontology selection. (iii) In Section 3.2.6 on page 89 we show how to learn an attribute-space distance metric to optimally weight attributes for re-identification, and do so in a synergistic way with low-level features. (iv) We evaluate our model in Section 3.3 and show significantly improved re-identification performance compared to conventional feature-based techniques on the two largest benchmark datasets. In the subsequent sections, we provide additional analysis and insight into the results, including contrast against zero-shot re-identification from attribute-profile descriptions.

3.2 Computing Attributes for Re-identification

3.2.1 Ontology Selection

The majority of recent work on attributes looks to human expertise in answer to the question as to which attributes to learn. Typically, ontology selection is performed manually prior to research

or via learning from existing metadata [18]. Recall from Section 2.2.1 on page 62 that hand-picked ontologies can be broadly categorised as top-down and bottom-up. In the top-down case, ontology selection may be predicated on the knowledge of experienced human domain-experts. In the latter it may be based on the intuition of vision researchers, based on factors such as how detectable an attribute might be with available methods or data availability.

For the purpose of automated re-identification, we are concerned with descriptions that permit us reliably discriminate; that is to say we wish to eliminate identity ambiguity between individuals. Ontology selection therefore is guided by two factors: *computability* and *usefulness*. That is, *detectable* attributes, which can be detected reliably using current machine learning methods and available data [58], and *discriminative* (informative) attributes which, if known, would allow people to be effectively disambiguated [124].

The notion of discriminative attributes encompasses a nuance. Humans share a vast prior pool of potential attributes and experience. If required to describe a person in a way which uniquely identifies them against a gallery of alternatives, they typically choose a short description in terms of the rare attributes which uniquely discriminate the target individual (e.g. "imperial moustache"). In contrast, in the ideal discriminative ontology of attributes for automated processing, each attribute should be uncorrelated with all others, and should occur in exactly half of the population (e.g. male versus female). In this way no one attribute can distinguish a person uniquely, but together they effectively disambiguate the population: a "binary search" strategy. There are two reasons for this: constraining the ontology size, and training data requirement.

Ontology size: Given a "binary search" ontology, any individual can be uniquely identified among a population of n candidates with only an $O(\log(n))$ sized attribute ontology or description. In contrast, the single rare-attribute strategy favoured by people means that while a person may be identified with a short length 1 attribute description, an ontology size and computation size $O(n)$ may be required to describe, interpret and identify this person.

Training data: We employ individual "binary" classifiers to model our ontology, thus each training image may be re-used and be (equally) informative for all n attributes (attributes are typically positive for half the images). In contrast, the single rare-attribute strategy would require an infeasible n times as much training data, because different data would be needed for each attribute (e.g. finding a significant number of wearers of imperial moustaches) to train the detectors. In practice, rare attributes do not have enough training data to learn good classifiers, and are thus not reliably

detectable. A final consideration is the visual subtlety of the attributes, which humans may be able to easily pick out based on their lifetime of experience but which would require prohibitive amounts of training data as well as feature/classifier engineering for machines to detect.

Whether or not a particular ontology is detectable and discriminative cannot therefore be evaluated prior to examination of representative data. However, given a putative ontology and a representative and annotated training set, the detectability of the ontology can be measured by the test performance of the trained detectors whilst the discriminativeness of the ontology can be measured by the mutual information (MI) between the attributes and person identity. The question of how to trade off discriminativeness and detectability when selecting an ontology on the basis of maximum predicted performance is not completely clear [101, 102]. However, we will take some steps to address this issue in Section 3.2.6 on page 89.

Figure 3.1: Positive instances of our ontology from (top) the VIPeR and (bottom) the PRID datasets.

redshirt	blueshirt	lightshirt
darkshirt	greenshirt	nocoats
notlightdarkjeanscolour	darkbottoms	lightbottoms
hassatchel	barelegs	shorts
jeans	male	skirt
patterned	midhair	darkhair
bald	hashandbagcarrierbag	hasbackpack

Table 3.1: Our attribute ontology for re-identification.

3.2.2 Ontology Creation and Data Annotation

Given the considerations discussed in the previous section, we select our ontology jointly based on four criteria. (i) We are informed by the operational procedures of human experts [132] as well as (ii) prioritising suitable findings from [176, 101, 102, 153], (iii) whether the ontology is favourably distributed in the data (binary search) and (iv) those which are likely to be detectable (sufficient training data and avoiding subtlety).

Specifically, we define the following space of $N_a = 21$ binary attributes (Table 3.1 on the preceding page). Ten of these attributes are related to colour, one to texture and the remaining ten are related to soft biometrics. Figure 3.1 on the previous page shows a visual example of each attribute.

Human annotation of attribute labels is costly in terms of both time and human effort. Due to the semantic nature of the attributes, accurate labelling can be especially challenging for cases where visual data can be impoverished. Typically problems can arise where (i) ontology definition allows for ambiguity between members of the ontology, and (ii) boundary cases are difficult for an annotator to classify according to a binary system with confidence. These circumstances can be natural places for subjective labelling errors [161].

To investigate the significance of this issue, we independently double-annotated the Person Re-ID (PRID) dataset [75] for our attribute ontology. Figure 3.2 illustrates frequency of label disagreements for each attribute in the PRID dataset measured as the Hamming distance between all annotations for that attribute across the dataset:

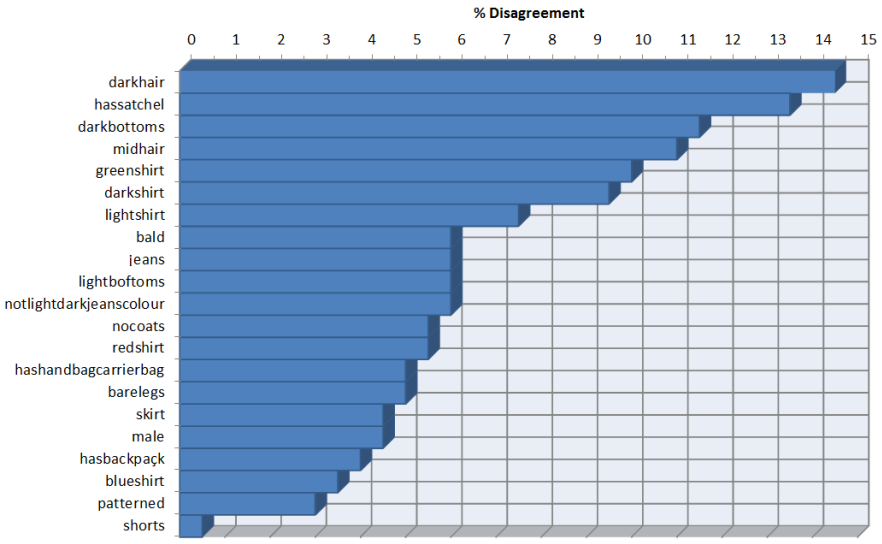


Figure 3.2: Annotation disagreement error frequencies for two annotators on PRID.

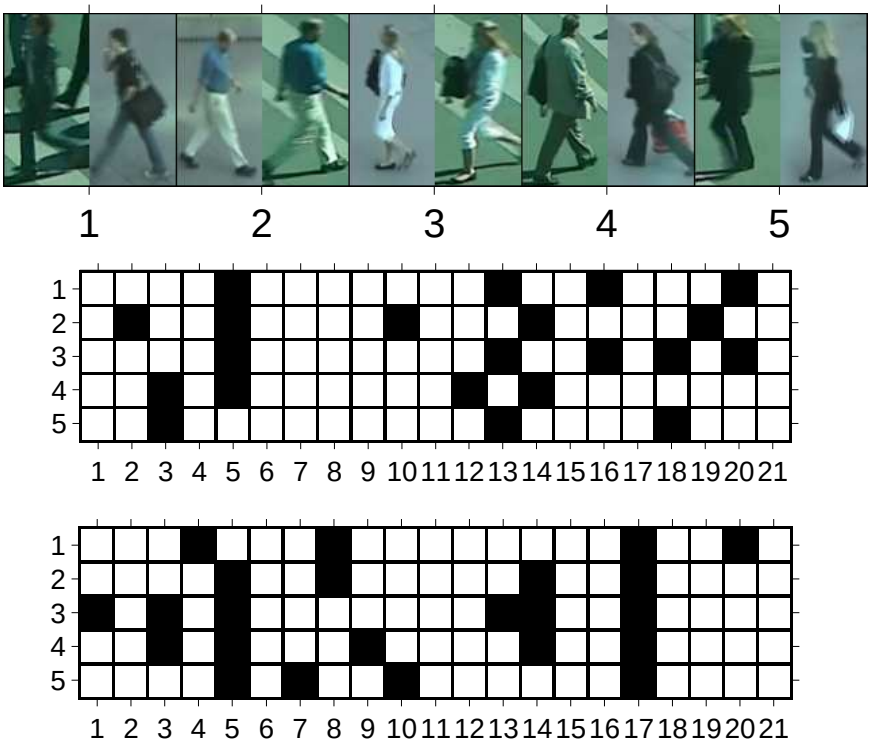


Figure 3.3: Top 5 pairs of pedestrian detections in PRID where annotators disagreed most (top row). Annotator #1's labels (middle), Annotator #2's labels (bottom). Each row is an attribute-profile for a pair of detections, columns are attributes and are arranged in the same order as Fig 3.2 on the facing page.

For attributes such as *shorts* or *gender*, uncertainty and therefore error is low. However, attributes whose boundary cases may be less well globally agreed upon can be considered to have the highest relative error between annotators. For example, in Figure 3.2 on the preceding page attributes *hassatchel* and *darkhair* are most disagreed upon since lighting variations make determining darkness of hair difficult in some instances and satchel refers to a wide variety of rigid or non-rigid containers held in multiple ways. This means that attributes such as *darkhair* and *hassatchel* may effectively be subject to a significant rate of label noise [194] in the training data and hence perform poorly. This adds another source of variability in reliability of attribute detection which will have to be accounted for later. Figure 3.3 illustrates pairs of individuals in the PRID dataset whose shared attribute-profiles were the most disagreed upon. The figure highlights the extent of noise that can be introduced through semantic labelling errors, a topic we will revisit later in Section 3.2.6 on page 89.

3.2.3 Feature Extraction

To detect attributes, we first select well-defined and informative low-level features with which to train robust classifiers. We wish to choose a feature which is also typically used for re-identification in order to enable later direct comparison between conventional and attribute-space re-identification in a way which controls for the input feature used. The descriptors we used for re-identification include the Symmetry Driven Accumulation of Local Features (SDALF) [47] and Ensemble of Localised Features (ELF) [68].

The content of our ontology includes semantic attributes such as jeans, shirt colours, gender. We can infer that the information necessary for humans to distinguish these items is present visually, and wish to select a feature that incorporates information pertaining to colour, texture and spatial information. For our purposes, SDALF fulfils the requirements for our ontology but does not produce positive semi-definite distances, therefore ruling it out for classification using kernel methods. Alternatively, we therefore exploit ELF.

To that end, we first extract an 2784-dimensional low-level colour and texture feature vector denoted $\mathbf{x}$ from each person image I following the method and parameters used in [147]. This consists of 464-dimensional feature vectors extracted from six equal sized horizontal strips from the image. Each strip uses 8 colour channels (RGB, HSV and YCbCr) and 21 texture filters (Gabor, Schmid) derived from the luminance channel. We use the same parameter choices for γ, λ, θ and σ^2 as proposed in [147] for Gabor filter extraction, and for τ and σ for Schmid extraction. Finally, we use a bin size of 16 to quantise each channel.

3.2.4 Attribute Detection

Classifier training and attribute feature construction

We train a Support Vector Machine (SVM) [155] for each attribute. We use Chang et al.'s LIBSVM [28] and investigate Linear, RBF, χ^2 and Intersection kernels. We select the intersection kernel as it compares closely with χ^2 but is faster to compute. Our experiments on LIBSVM performance vs. attribute training time show the intersection kernel as being a good combination of calculation time and accuracy. For example, training the attribute ontology results in 65.4% mean accuracy with 0.8 hours training for the intersection kernel, as compared to the χ^2 kernel (63.8% with 4.1 hours), the RBF kernel (65.9% with 0.76 hours and the linear kernel (61.8% with 1.2 hours) respectively with LIBSVM. Although RBF is computed slightly faster

and has similar accuracy, we select the intersection kernel overall, since an RBF kernel requires two parameters which would require additional cross-validation, and we can avoid this with the intersection kernel with little penalty. Providing LibSVM with pre-built kernels reduces training time considerably in all cases.

For each attribute, we perform cross validation to select values for SVM slack parameter C from the set $C \in \{-10, \ldots, 10\}$ with increments of $\varepsilon = 1$. The SVM scores are probability mapped, so each attribute detector i outputs a posterior $p(a_i|\mathbf{x})$. We follow the standard approach for mapping SVM scores to posterior probabilities [144] as implemented by libSVM [28].

Spatial Feature Selection

Since some attributes (e.g. shorts) are highly unlikely to appear outside of their expected spatial location, one might ask whether it is possible to improve performance by discriminatively selecting or weighting the individual strips within the feature vector (Section 3.2.3 on the facing page). We experimented with defining a kernel for each strip as well as for the entire image, and training multi-kernel learning SVM using the DOGMA online kernel learning library with *Online-Batch Strongly Convex mUlti keRnel lEarning* (Obscure) method as in [136, 46]. This approach discriminatively optimises the weight for each kernel in order to improve classifier performance and has been shown to improve performance when combining multiple features. However in this case it did not reliably improve on the conventional SVM approach, presumably due to the relatively sparse and imbalanced training data being insufficient to correctly tune the inter-kernel weight.

Imbalanced Attribute Training

The prevalence of each attribute in a given dataset tends to vary dramatically and some attributes have a limited number of positive examples in an absolute sense as a result. This imbalance can cause discriminative classifiers such as SVMs to produce biased or degenerate results. There are various popular approaches to dealing with imbalanced data [72], such as synthesising further examples from the minority class to improve the definition of the decision boundary, for example using SMOTE [29] or weighting SVM instances or mis-classification penalties [72, 2]. However, neither of these methods outperformed simple subsampling in our case.

To avoid bias due to imbalanced data, we therefore simply train each attribute detector with all the positive training examples of that attribute, and obtain the same number of negative examples by sub-sampling the rest of the data at regular intervals.

Mid-level Attribute Representation

Given the learned bank of attribute detectors, at test time we generate mid-level features as $1 \times N_a$ sized vectors of classification posteriors which we use to represent the probability that each attribute is present in the detection. Effectively we have projected the high dimensional, low-level features onto a mid-level, low-dimensional semantic attribute space. In particular, each person image is now represented in semantic attribute space by stacking the posteriors from each attribute detector into the N_a dimensional vector: $A(\mathbf{x}) = [p(a_1|\mathbf{x}_i), \ldots, p(a_{N_a}|\mathbf{x}_{N_a})]^T$.

3.2.5 Attribute Fusion with Low-level Features

The attribute representation, since it is trained using human expertise, encodes substantially different information to the LLFs used in it's training. Because of this, the trained attribute representation remains synergistic with and complementary to low-level features, meaning that we are able to fuse LLF representations with the attribute representation for better performance.

To use our attributes for re-identification, we can define a distance solely on the attribute space, or use the attribute distance in conjunction with conventional distance between low-level features such as SDALF [47] and ELF [67]. SDALF provides effective features for a non-learning nearest-neighbour (NN) approach while ELF has been widely used by model-based learning approaches [147, 187]. We also use it as the feature for our attribute detectors in Section 3.2.3 on page 86.

We therefore introduce a rather general formulation of a distance metric between two images I_p and I_g which combines both multiple attributes and multiple low-level features as follows:

$$d_{\mathbf{w}^L,\mathbf{w}^A}(I_p,I_g) = \sum_{l \in LL} w_l^L \, d_l^L \left(L_l(I_p), L_l(I_g)\right) + d_{\mathbf{w}^A}^A \left(A(I_p), A(I_g)\right). \tag{3.1}$$

Here Equation (3.1), the first term corresponds to the contribution from a set LL of low-level distance measures, where $L_l(I_p)$ denotes extraction of type l low-level features from image I_p, d_l^L denotes the distance metric defined for low-level feature type l, and w_l^L is a weighting factor for each feature type l. Equation (3.1) (second term) corresponds to the contribution from our attribute-based distance metrics. Where $A(I_p)$ denotes the attribute encoding of image I_p. For the attribute-space distance we experiment with two metrics: weighted L1 (Equation 3.2 on the facing page) and weighted Euclidean (Equation 3.3 on the next page).

$$d^A_{\mathbf{w}^A}(I_p, I_g) = (\mathbf{w}^A)^T \left| (A(\mathbf{x}_p) - A(\mathbf{x}_g)) \right|, \tag{3.2}$$

$$d^A_{\mathbf{w}^A}(I_p, I_g) = \sqrt{\sum_i w^A_i \left(p(a_i|\mathbf{x}_{p,i}) - p(a_i|\mathbf{x}_{g,i}) \right)^2}. \tag{3.3}$$

3.2.6 Attribute Selection and Weighting

As discussed earlier, all attributes are not equal due to variability in how reliably they are measured due to imbalance, subtlety (detectability) and how informative they are about identity (discriminability). How to account for variable detectability and discriminability of each attribute ($\mathbf{w}^A$), and how to weight attributes relative to low-level features ($\mathbf{w}^{LL}$) are important challenges which we discuss now.

Exhaustively searching the N_a dimensional space of weights directly to determine attribute selection and weighting is computationally intractable. However, we can re-formulate the re-identification task as an optimisation problem and apply standard optimisation methods [131] to search for a good configuration of weights.

Importantly, we only search $|\mathbf{w}^A| = N_a = 21$ parameters for the within-attribute-space metric $d^A_{\mathbf{w}^A}(\cdot,\cdot)$. and one or two parameters for weighting attributes relative to low-level features. In contrast to previous learners for low-level features [147, 188, 192] which must optimise hundreds or thousands of parameters, this gives us considerable flexibility in terms of computation requirement of the objective.

An interesting question is therefore what is the ideal criterion for optimisation. Previous studies have considered optimising, e.g. relative rank [147] and relative distance [192, 75]. While effective, these metrics are indirect proxies for what the re-identification application ultimately cares about, which is the average rank of the true match to a probe within the gallery set, which we call Expected Rank (ER). That is, how far does the operator have to look down the list before finding the target. See Section 3.3 for more discussion.

We introduce the following objective for expected rank:

$$ER = \frac{1}{|P|} \sum_{p \in P} \sum_{g \in G} L_{\mathbf{w}}(D_{pp}, D_{pg}) + \lambda \parallel \mathbf{w} - \mathbf{w}_0 \parallel, \tag{3.4}$$

where D_{pg} is the matrix of distances, from probe image p to gallery image g; L_w is a loss function which can penalise the objective according to the relative distance of the true match D_{pp} versus false matches D_{pg}; and $\mathbf{w}_0$ is a regulariser bias with strength λ. To complete the definition of the objective, we define the loss function L as in Equation (3.5 on the following page) where $\mathbf{I}$ is an

indicator function which returns 1 when the parameter is true. That is, imposing a penalty every time a false match is ranked ahead of the true match. The overall objective, Equation (3.4 on the previous page) thus returns the expected rank of the true match. This is now a good objective, because it directly reflects the relevant end-user metric for effectiveness of the system. However it is hard to efficiently optimise because it is non-smooth: a small change to the weights $\mathbf{w}$ may have exactly zero change to the expected rank (the optimisation surface is piece-wise linear). We therefore soften this loss-function using a sigmoid, as in Equation (3.6), which is now smooth and differentiable. This finally allows efficient gradient-based optimisation with Newton [114] or conjugate-gradient methods [131].

$$L_{\mathbf{w}}^{HardRank,ER} = \mathbf{I}(d_{pp} - d_{pg} > 0). \tag{3.5}$$

$$L_{\mathbf{w}}^{Sigmoid,ER} = \sigma(d_{pp} - d_{pg}). \tag{3.6}$$

We initialise $\mathbf{w}_{initial} = 1$. To prevent over fitting, we use regularisation parameters $\mathbf{w}_0$=1, and $\lambda = 0.2$ (i.e., all weights are assumed to be equally important *a priori*) and set the sigmoid scale to $k = 32$. Finally we perform fusion with low-level features, Equation 3.1 on page 88, using both SDALF and ELF.

In summary, this process uses gradient-descent to search for a setting of weights $\mathbf{w}$ for each LLF and for each attribute, Equation (3.1 on page 88) that will (locally) minimise the expected rank within the gallery of the true match to each probe image, Equation (3.4 on the previous page). See Algorithm 1 on the facing page for an overview of our complete system.

3.3 Experiments

3.3.1 Datasets

We select two challenging datasets with which to validate our model, the Viewpoint Invariant Pedestrian Recognition dataset (VIPeR) [67] and PRID [75]. VIPeR contains 632 pedestrian image pairs from two cameras with different viewpoint, pose and lighting. Images are scaled to 128x48 pixels. We follow [67, 47] in considering Cam B as the gallery set and Cam A as the probe set. Performance is evaluated by matching each test image in Cam A against the Cam B gallery.

PRID is provided as both multi-shot and single-shot data. It consists of two camera views overlooking an urban environment from a distance and from fixed viewpoints. As a result PRID features low pose variability with the majority of people captured in profile. The first 200 shots

Algorithm 1 Attributes-based Re-identification

Training

for each Attribute **do**

Subsample majority class to length of minority class

Cross-validate to obtain parameter C that gives best average accuracy.

Retrain SVM on all training data with selected C

end for

Determine inter and intra-attribute weighting $\mathbf{w}$ by minimising Equation (3.4 on page 89).

Testing (Re-identification)

for each Person $\mathbf{x}_g \in$ gallery set **do**

Classify each attribute a

Stack attribute posteriors into person signature $A(\mathbf{x}_g)$.

end for

for each Person $\mathbf{x}_p \in$ probe set **do**

Classify each attribute a

Stack attribute posteriors into person signature $A(\mathbf{x}_p)$.

Compute distance to gallery set fusing attribute and LLF cues with weight $\mathbf{w}$. (Equation (3.1 on page 88))

Nearest-neighbour re-identification in gallery according to their similarity to person $\mathbf{x}_p$.

end for

in each view correspond to the same person, however the remaining shots only appear once in the dataset. To maximise comparability with VIPeR, we use the single-shot version and use the first 200 shots from each view. Images are scaled to 128x64 pixels.

For each dataset, we divide the available data into training, validation and test partitions. We initially train classifiers and produce attribute representations from the training portion, and then optimise the attribute weighting as described in Section 3.2.6 on page 89 using the validation set. We then retrain the classifiers on both the training and validation portions, while re-identification performance is reported on the held out test portion.

We quantify re-identification performance using three standard metrics and one less common

metric. The standard re-identification metrics are performance at rank n, cumulative matching characteristic (CMC) curves, and normalised area under the CMC curve [67, 47]. Performance at rank n reports the probability that the correct match occurs within the first n ranked results from the gallery. The CMC curve plots this value for all n, and the nAUC summarises the area under the CMC curve (so perfect nAUC is 1.0 and chance nAUC is 0.5).

We additionally report Expected Rank (ER), as advocated by Avraham et al. [6] as CMC Expectation. The ER reflects the mean rank of the true matches and is a useful statistic for our purposes; in contrast to the standard metrics, lower ER scores are more desirable and indicate that on average the correct matches are distributed more toward the lower ranks. (So perfect ER is 1 and random ER would be half the gallery size). In particular ER has the advantage of a highly relevant practical interpretation: it is the average number of returned images the operator will have to scan before reaching the true match.

We compare the following re-identification methods: (1) SDALF [47] using code provided by the authors (note that SDALF is already shown to decisively outperform [68]); (2) ELF: Prosser et al.'s [147] spatial variant of Ensemble of Localised Features (ELF) [67] using Strips of ELF; (3) Attributes: Raw attribute based re-identification (Euclidean distance); (4) **OAR**: our Optimised Attribute based Re-identification method with weighting between low-level features and within attributes learned by directly minimising the Expected Rank (Section 3.2.6 on page 89).

3.3.2 Attribute Analysis

We first analyse the intrinsic discriminative potential of our attribute ontology independently of how reliably detectable the attributes are (assuming perfect detectability). This analysis provides an upper bound of performance that would be obtainable with sufficiently advanced attribute detectors. Figure 3.6 on page 95 reports the prevalence of each attribute in the datasets. Many attributes have prevalence near to 50%, which is reflected in their higher mutual information with person identity. As we discussed earlier this is a desirable property because it means each additional attribute known can potentially halve the number of possible matches. Whether this is realised or not depends on if attributes are correlated/redundant, in which case each additional redundant attribute provides less marginal benefit. To check this we compute the correlation coefficient between all attributes, and found that the average inter-attribute correlation was only 0.07. We therefore expect the attribute ontology to be effective.

Figure 3.4 on the facing page shows a histogram summarising how many people are uniquely

identifiable solely using attributes and how many would be confused to a greater or lesser extent. The peak around unique/unambiguous shows that a clear majority of people can be uniquely or otherwise near-uniquely identified by their attribute-profile alone, while the tail shows that there are a small number of people with very generic profiles. This observation is important; near-uniqueness means that approaches which rank distances between attribute-profiles are still likely to feature the correct match high enough in the ranked list to be of use to human operators.

The CMC curve (for gallery size p=632) that would be obtained assuming perfect attribute classifiers is shown in Figure 3.5 on page 95. This impressive result (nAUC near a perfect score of 1.0) highlights the potential for attribute-based re-identification. Also shown are the results with only the top 5 or 10 attributes (sorted by mutual information with identity), and a random 10 attributes. This shows that: (i) as few as 10 attributes are sufficient if they are *informative* (i.e. high MI) and perfectly detectable, while 5 is too few; and (ii) attributes with high MI are significantly more useful than low MI (always present or absent) attributes.

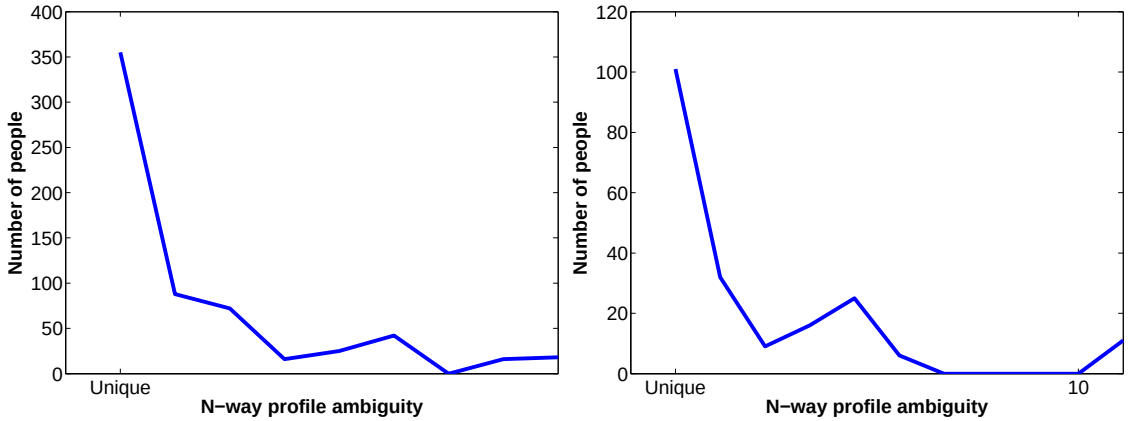


Figure 3.4: Uniqueness of attribute descriptions in a population, (i) VIPeR and (ii) PRID. The peak around unique shows that most people are uniquely identifiable by attributes.

3.3.3 Attribute Detection

Given the analysis of the intrinsic effectiveness of the ontology in the previous section, the next question is whether the selected attributes can indeed be detected or not. Attribute detection on both VIPeR and PRID achieves reasonable levels on both balanced and unbalanced datasets as seen in Table 3.2 on the following page. (dash indicates failure to train due to insufficient data). For all datasets, a minimum of 9 classifiers can be trained on unbalanced PRID, and 16 on unbalanced VIPeR, in both cases some attribute classifiers are unable to train due to extreme

	VIPeR (u)	VIPeR (b)	PRID (u)	PRID (b)
redshirt	79.6	80.9	–	41.3
blueshirt	62.7	68.3	–	59.6
lightshirt	80.6	82.2	81.6	80.6
darkshirt	82.2	84.0	79.0	79.5
greenshirt	57.3	72.1	–	–
nocoats	68.5	69.7	–	31.3
notlightdarkjeanscolour	57.6	69.1	–	–
darkbottoms	74.4	75.0	72.2	67.3
lightbottoms	75.3	74.7	76.0	74.0
hassatchel	–	56.0	51.9	55.0
barelegs	60.4	74.4	–	50.2
shorts	53.1	76.1	–	–
jeans	73.6	78.0	57.1	69.4
male	66.7	68.0	52.1	54.0
skirt	–	68.8	–	44.6
patterned	–	60.8	–	–
midhair	55.2	64.6	69.4	70.4
darkhair	60.0	60.0	75.4	75.4
bald	–	–	–	40.2
hashandbagcarrierbag	–	54.5	–	59.4
hasbackpack	63.4	68.6	–	48.3
Mean	66.9	70.3	68.3	66.2

Table 3.2: Attribute Classifier training and test accuracies (%) for VIPeR and PRID, for both the balanced (b) and unbalanced (ub) datasets.

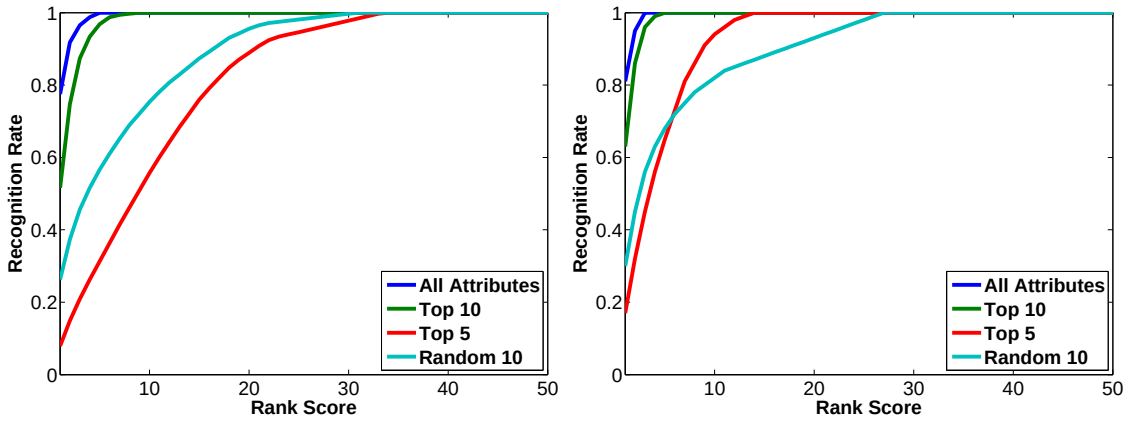


Figure 3.5: Best-case (assuming perfect attribute detection) re-identification using attributes with highest n ground-truth Mutual Information scores, (i) VIPeR and (ii) PRID.

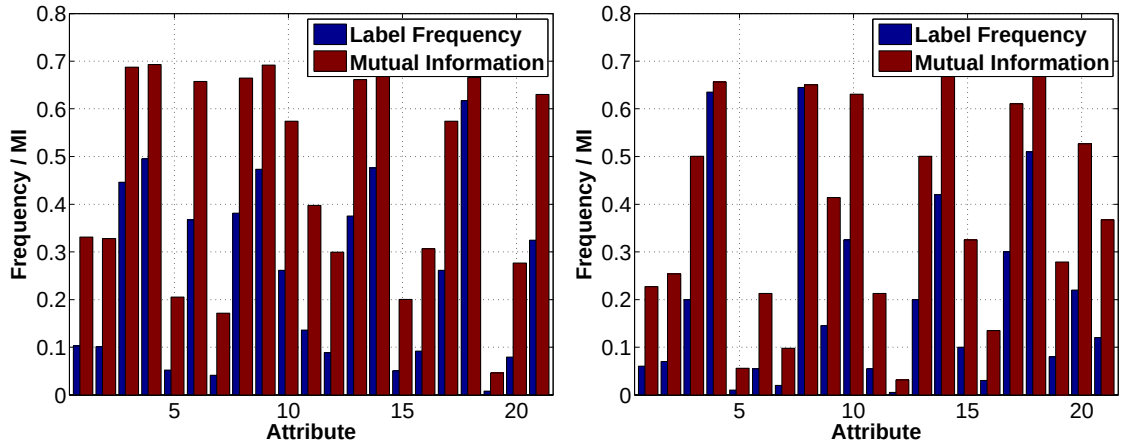


Figure 3.6: Attribute occurrence frequencies and Attribute Mutual Information (MI) scores in VIPeR (left) and PRID (right).

class imbalances or data sparsity. Average accuracies for these datasets are also reasonable; 66.9% and 68.3% respectively. The benefit of sub-sampling negative data for attribute learning is highlighted in the improvement in the balanced datasets. Balancing in this case increases the number of successfully trained classifiers to 20 for balanced VIPeR and 16 on balanced PRID with mean accuracies rising to 70.3% for VIPeR. Balancing slightly reduces classification performance on PRID to an average of 66.2%.

3.3.4 Using Attributes to Re-identify

Given the previous analysis of discriminability and detectability of the attributes, we now address the central question of attributes for re-identification. We first consider the "raw" attribute re-identification case (i.e. no weighting or fusion; $\mathbf{w}^L = 0, \mathbf{w}_a = 1$ in Equation (3.1 on page 88)). The

re-identification performance of attributes alone is summarised in Table 3.3 in terms of expected rank. There are a few interesting points to note: (i) In most cases using $L2$ NN matching provides lower ER scores than $L1$ NN matching. (ii) On VIPeR and PRID, SDALF outperforms the other low-level features, and outperforms our basic attributes in VIPeR. (iii) Although the attribute-centric re-identification uses the *same low-level input features* (ELF), and the same L1/L2 NN matching strategy, attributes decisively outperform raw ELF. We can verify that this large difference is due to the semantic attribute space rather than the implicit dimensionality reduction effect of attributes by performing Principle Components Analysis (PCA) on ELF to reduce its dimensionality to the same as our attribute space ($N_a = 21$). In this case the re-identification performance is still significantly worse than the attribute-centric approach (See Table 3.3). The improvement over raw ELF is thus due to the attribute-centric approach.

VIPeR	$L1$	$L2$
ELF [147]	84.3	72.1
ELF PCA	85.3	74.5
Raw Attributes	**34.4**	**37.8**
SDALF [47]	44.0	
Chance Level	158	

PRID	$L1$	$L2$
ELF	28.2	37.0
ELF PCA	32.7	38.1
Raw Attributes	**24.1**	**24.4**
SDALF [47]	31.8	
Chance Level	50	

Table 3.3: Re-identification performance, we report Expected Rank (average rank of the true match) scores for VIPeR (left, gallery size $p = 316$) and PRID (right, gallery size $p = 100$) and compare different features and distance measures against our balanced attribute-features prior to fusion and weight selection. Smaller values indicate better re-identification performance.

3.3.5 Re-identification With Optimised Attributes

Given the promising results for vanilla attribute re-identification in the previous section, we finally investigate whether our complete model (including discriminative optimisation of weights to improve expected rank) can further improve performance. Figure 3.7 on the facing page and Table 3.4 on page 98 summarise final re-identification performance. In each case, optimising the attributes with the distance metric and fusing with low-level SDALF and ELF improves re-identification uniformly compared to using attributes or low-level features alone. Our approach improves ER by 38.3% and 35% on VIPeR, and 38.8% and 46.5% on PRID for the balanced and unbalanced cases vs SDALF and 66.9%, 65.1%, 77.1% and 80% vs ELF features.

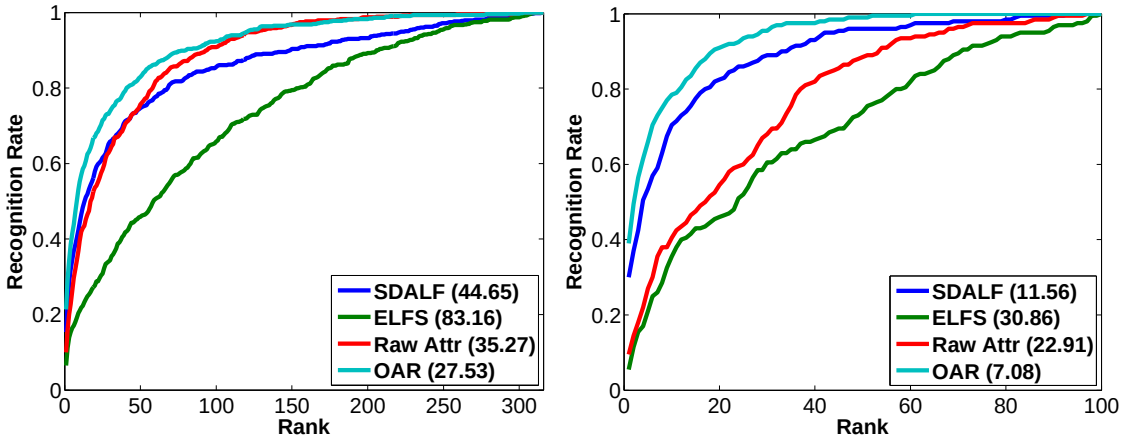


Figure 3.7: Final attribute re-identification CMC plots for (i) VIPeR and (ii) PRID, Gallery sizes $p = 316, p = 100$. Expected Rank is given in parentheses.

Critically for re-identification scenarios, the most important rank 1 accuracies are improved convincingly. For VIPeR, OAR improves 40% over SDALF in the balanced case, and 33.3% for unbalanced data. For PRID, OAR improves by 30% and 36.6%. As in the case of ER, rank is uniformly improved, indicating the increased likelihood that correct matches appear more frequently at earlier ranks using our approach.

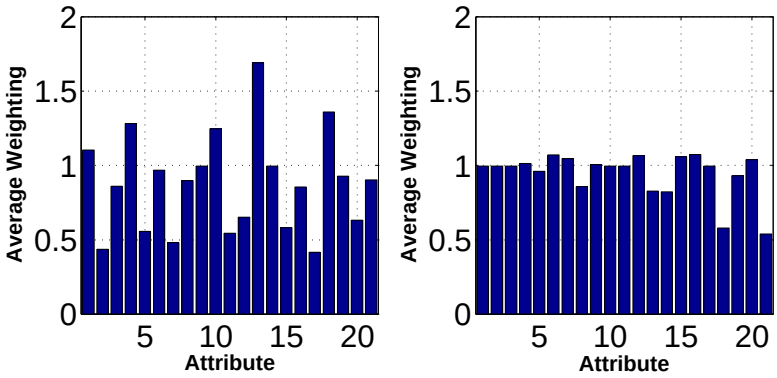


Figure 3.8: Final attribute feature weights for VIPeR (left) and PRID (right).

The learned weights for fusion between our attributes and low-level features indicate that SDALF is informative and useful for re-identification on both datasets. In contrast, ELF is substantially down-weighted to 18% compared to SDALF on PRID and on VIPeR, disabled entirely. This makes sense because SDALF is at least twice as effective as ELF for VIPeR (Table 3.3 on the facing page).

The intra-attribute weights (Figure 3.8) are relatively even on PRID but more varied on VIPeR where the highest weighted attributes (*jeans, hasbackpack, nocoats, midhair, shorts*) are

VIPeR	ER	Rank 1	Rank 5	Rank10	Rank25	nAUC
Farenzena et al. [47]	44.7	15.3	34.5	44.3	61.6	0.86
Prosser et al. [147]	83.2	6.5	16.5	21.0	30.9	0.74
Raw Attributes (b)	35.3	10.0	26.3	39.6	58.4	0.89
OAR (b)	**27.5**	**21.4**	**41.5**	**55.2**	**71.5**	**0.94**
Raw Attributes (u)	40.4	6.5	23.9	34.8	55.9	0.88
OAR (u)	29.0	19.6	39.7	54.1	71.2	0.91
PRID	ER	Rank 1	Rank 5	Rank10	Rank25	nAUC
Farenzena et al.	11.6	30.0	53.5	70.5	86.0	0.89
Prosser et al.	30.9	5.5	21.0	35.5	52.0	0.70
Raw Attributes (b)	22.9	9.5	27.0	40.5	60.0	0.78
OAR (b)	7.1	39.0	66.0	78.5	93.5	0.93
Raw Attributes (u)	20.8	8.5	28.5	44.0	69.0	0.80
OAR (u)	**6.2**	**41.5**	**69.0**	**82.5**	**95.0**	**0.95**

Table 3.4: Final attribute re-identification performance. We report Expected Rank scores [6] (lower scores indicate that overall, an operator will find the correct match appears lower down the ranks), Cumulative Match Characteristic (CMC) and normalised Area-Under-Curve (nAUC) scores (higher is better, the maximum nAUC score is 1). We further report accuracies for our approach using unbalanced data for comparison.

weighted at 1.43, 1.20, 1.17, 1.10 and 1.1; while the least informative attributes are *barelegs, lightshirt, greenshirt, patterned* and *hassatchel* which are weighted to 0.7, 0.7, 0.66, 0.65 and 0.75. Jeans is one of the attributes that is detected most accurately and is most common in the datasets, so its weight is expected to be high. However the others are more surprising, with some of the most accurate attributes such as *darkshirt* and *lightshirt* weighted relatively low (0.85 and 0.7). For PRID, *darkshirt, skirt, lightbottoms, lightshirt* and *darkbottoms* are most informative (1.19, 1.04, 1.02 and 1.03); *darkhair, midhair, bald, jeans* are the least (0.78, 0.8, 0.92, 0.86).

Interestingly, the most familiar indicators which might be expected to differentiate good versus bad attributes are not reflected in the final weighting. Classification accuracy, annotation error (label noise) and mutual information are not significantly correlated with the final weighting, meaning that some unreliably detectable and rare/low MI attributes actually turn out to be *useful* for re-identification with low expected rank; and vice-versa. Moreover, some of the

VIPeR	Rank 1	Rank 10	Rank 20	Rank 50	nAUC
OAR	21.4	55.2	71.5	82.9	**0.92**
Hirzer et al.[76]	22.0	63.0	78.0	93.0	-
Farenzena et al.[47]	9.7	31.7	46.5	66.6	0.82
Hirzer et al.[77]	**27.0**	**69.0**	**83.0**	**95.0**	-
Avraham et al.[6]	15.9	59.7	78.3	-	-
Zheng et al.[188, 192]	15.7	53.9	70.1	-	-
Prosser et al.[147]	14.6	50.9	66.8	-	-

Table 3.5: Comparison of results between our OAR method (Optimised Attribute Re-identification) and other state of art results for the VIPeR dataset.

weightings vary dramatically between dataset, for example, the attribute *jeans* is the strongest weighted attribute on VIPeR, however it is one of the lowest on PRID despite being reasonably accurate and prevalent on both datasets. These two observations both show (i) the necessity of jointly learning a combined weighting for all the attributes, (ii) doing so with a relevant objective function (such as ER), and (iii) learning a model which is adapted for the statistics of each given dataset/scenario.

In Table 3.5, we compare our approach with the performance other methods as reported in their evaluations. In this case the cross-validation folds are not the same, so the results are not exactly comparable, however they should be indicative. Our approach performs comparably to [76] and convincingly compared to [47, 188, 192] and [147]. Both [77] and [6] exploit pairwise learning; in [6] a binary classifier is trained on correct and incorrect pairs of detections in order to learn the projection from one camera to another, in [77] incorrect (i.e., matches that are nearer to the probe than the true match) detections are directly mapped further away whilst similar but correct matches are mapped closer together. Our approach is eventually outperformed by [77], however [77] learns a full covariance distance matrix in contrast to our simple diagonal matrix, and despite this we remain reasonably competitive.

3.3.6 Zero-shot Identification

In Section 3.3.2 on page 92 we showed that with perfect attribute detections, highly accurate re-identification is possible. Even with a mere 10 attributes, near-perfect re-identification can

be performed. Zero-shot identification is the task of generating an attribute-profile either manually or from a different modality of data, then matching individuals in the gallery set via their attributes. This is highly topical for surveillance: consider the case where a suspect is escaping through a public area surveilled by CCTV. The authorities in this situation may have enough information build a semantic-attribute-profile of the suspect using attributes taken from eyewitness descriptions.

In zero-shot identification (a special case of re-identification) we replace the probe image with a manually specified attribute description. To test this problem setting, we match the ground truth attribute-profiles of probe persons against their inferred attribute-profiles in the gallery as in [174].

An interesting question one might ask is whether this is expected to be better or worse than conventional attribute-space re-identification based on attributes detected from a probe *image*. One might expect zero-shot performance to be better because we know that in the absence of noise, attribute re-identification performs admirably (Section 3.3.2 on page 92 and Figure 3.5 on page 95) – and there are two sources of noise (attribute detection inaccuracies in the probe and target images) of which the former noise source has been removed in the zero-shot case. In this case, a man-in-the-loop approach to querying might be desirable, even if a probe image is available. That is, the operator could quickly indicate the ground-truth attributes for the probe image and search based on this (noise-free) ground-truth.

Table 3.6 on the facing page shows re-identification performance for both datasets. Surprisingly, while the performance is encouraging, it is not as compelling as when the profile is constructed by our classifiers, *despite the elimination of the noise on the probe images*.

This significant difference between the zero-shot case we outline here and the conventional case we discuss in the previous section turns out to be because of *noise correlation*. Intuitively, consider that if someone with a hard-to-classify hairstyle is classified in one camera with some error ($p(a_{hair}|\mathbf{x}) - a^{true_hair}$), then this person might also be classified in another camera with an error *in the same direction*. In this case, using the ground-truth attribute in one camera will actually be detrimental to re-identification performance.

To verify this explanation, we perform Pearson's product-moment correlation analysis [143] on the error (difference between ground-truth labels and the predicted attributes) between the probe and gallery sets. The average cross-camera error correlation coefficient is 0.93 in VIPeR

and 0.97 in PRID, and all of the correlation coefficients were statistically significant ($p < 0.05$).

Although these results show that man-in-the-loop zero-shot identification - if intended to replace a probe image - may not always be beneficial, it is still evident that zero-shot performs reasonably in general and is a valuable capability for the case where descriptions are verbal rather than extracted from a visual example.

	ExpRank	Rank 1	Rank 5	Rank10	Rank25
VIPER (u)	50.1	6.0	17.1	26.0	48.1
VIPER (b)	54.8	5.4	14.9	25.3	44.9
PRID (u)	19.2	8.0	29.0	47.0	73.0
PRID (b)	26.1	3.0	16.0	32.0	62.0

Table 3.6: Zero-shot re-identification results for VIPeR and PRID.

3.4 Discussion

In this chapter, we have shown how mid-level attributes trained using semantic cues from human experts [132] can be an effective representation for re-identification and (zero-shot) identification. Moreover, this provides a different modality to standard low-level features and thus synergistic opportunities for fusion.

Existing approaches to re-identification [47, 147, 67] focus on high-dimensional low-level features which aim to be discriminative for identity yet invariant to view and lighting. However, these variance and invariance properties are hard to obtain simultaneously, thus limiting such features effectiveness for re-identification. In contrast, attributes provide a low-dimensional mid-level representation which are discriminative by construction (see Section 3.2.1 on page 81) and make no strong view invariance assumptions (variability in appearance of each attribute is learned by the classifier with sufficient training data).

Importantly, although individual attributes vary in robustness and informativeness, attributes provide a strong cue for identity. Their low-dimensional nature means they are also amenable to discriminatively learning a good distance metric, in contrast to the challenging optimisation required for high-dimensional LLFs [188, 192]. In developing a separate cue-modality, our approach is potentially complementary to the majority of existing approaches, whether focused on low-level features [47], or learning methods [188, 192]. Although the representation we in-

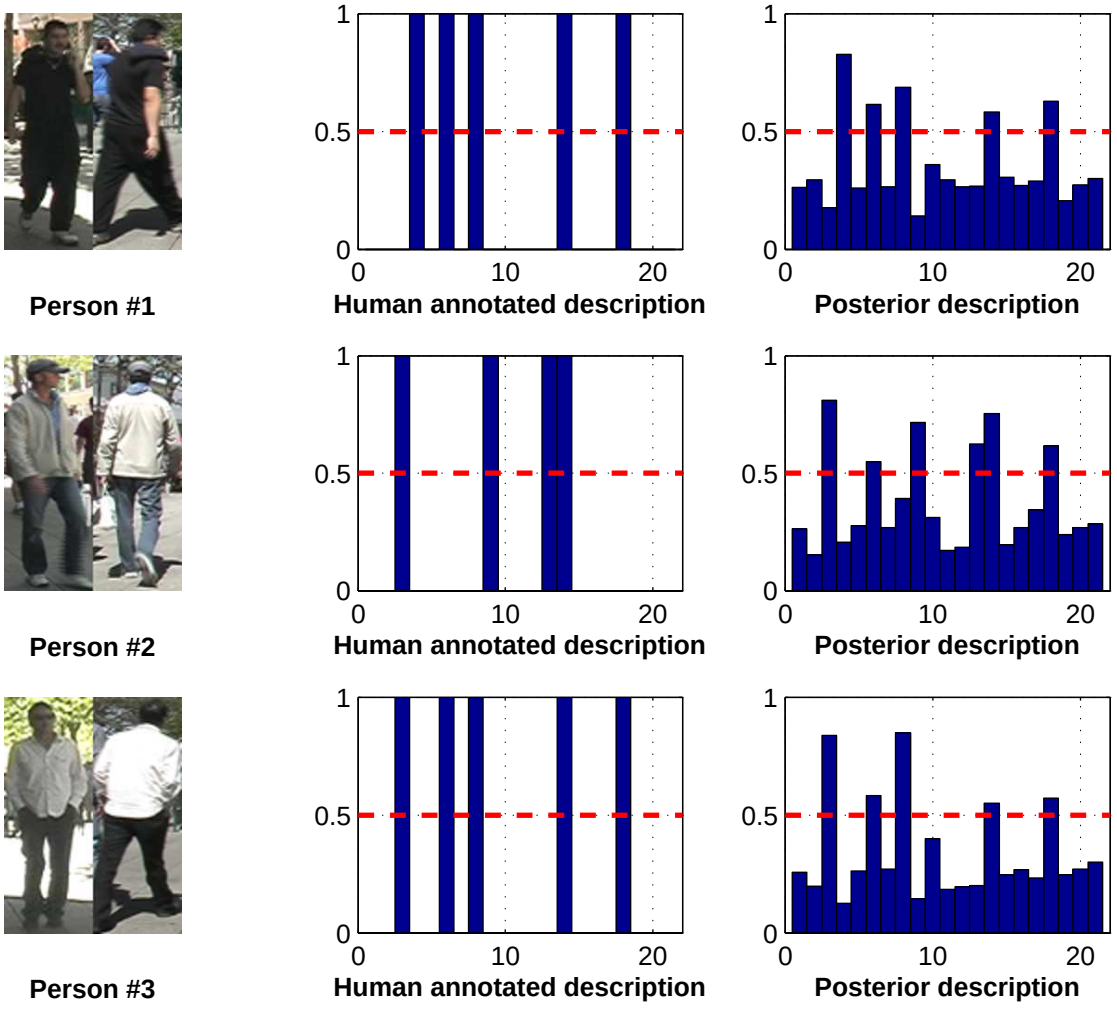


Figure 3.9: Success cases for Zero-shot re-identification on VIPeR. The left column shows two probe images; i) is the image annotated by a human operator and ii) is the correct rank #1 match as selected by our zero-shot re-identification system. The human-annotated probe descriptions (middle) and the matched attribute-feature gallery descriptions (right) are notably similar for each person; the attribute detections from the gallery closely resemble the human-annotated attributes (particularly those above red line).

troduce in this chapter has excellent potential, it requires significant human effort to label the data required to train each attribute classifier and for each new camera. In the next chapter, we examine how to generate a similar representation in a more scalable manner.

Chapter 4

Hunting Attributes in the Wild

In this chapter, we show how to automatically discover attributes that provide a valuable representation which significantly improves re-identification performance on a variety of challenging datasets. Existing attribute representations do not generalise across camera deployments. Thus, this standard strategy currently requires the prohibitive effort of annotating a vector of person attributes for each individual in a large training set – for each given deployment/dataset. In this chapter we take a different approach and automatically discover a semantic attribute ontology, and learn an effective associated representation by crawling large volumes of Internet data. In addition to eliminating the necessity for per-dataset annotation, by training on a much larger and more diverse array of examples this representation is more view-invariant and generalisable than attributes trained at conventional small scales.

4.1 Problem Definition

Feature-centric approaches to improving re-identification [47] typically suffer from the problem of it being extremely challenging to create features that are more than just weakly able to distinguish people reliably, whilst simultaneously still being invariant to all the practical visual covariates such as motion blur, clutter, view angle and pose change, lighting and occlusion. In contrast, learning re-identification models that discriminatively maximise re-identification performance, for example metric learning [77] and support vector machines (SVM) [147, 6] typically require copious human annotation and high quantities of data. These lines of inquiry are nevertheless synergistic because better feature representations tend to improve a given discrim-

inative algorithm applied downstream, while applying better discriminative methods to a given representation also tends to improve results.

The recent line of work [103, 117, 153, 174] in feature/representation learning draws inspiration from the practices of human experts. Human operators focus their attention on noting and matching distinct semantic characteristics, or *attributes*, to simplify their task. These may correspond to distinct soft-biometric, appearance or functional properties such as gender or clothing-style. Attribute-centric approaches learn a low-dimensional feature representation that corresponds to such semantic properties. They typically approach this by asking expert operators to define an ontology of such characteristics, collecting and annotating site specific training data with a vector of attributes per person, or training computer vision models to detect attributes. Then, the estimated attributes of each person can be taken as a representation for re-identification. However, this top-down human-defined attribute approach has some critical limitations: (i) It requires costly attribute annotation of scene-specific training data. This is significantly more costly than person-identity information used to train discriminative matching models. (ii) Top-down definition of attributes does not guarantee that they are visually computable by computer vision techniques given visual surveillance data. (iii) Due to the limited scalability of the annotation approach, the annotated data are likely to be too small scale to learn accurate and robust detectors for each attribute of interest.

Thus far, the reader can be forgiven for thinking our motivation for this chapter is more or less identical to the motivation for Chapter 3. However, we note in Chapter 3 that attributes trained discriminatively from real-world surveillance data depend on (i) data volume, (ii) classifier accuracy, (iii) class imbalance, and (iv) the availability and quality of human-expert defined labels – on the target data. This chapter will specifically consider (i) and (iv), which present significant challenges for representation learning due to the cost involved in acquiring fresh labels and data from human experts or real-world scenes.

4.1.1 Hunting Attributes for Re-identification

In the following sections we address these issues by taking a very different data-driven [30, 126] approach to learning attributes for re-identification rather than learning them directly as in Chapter 3. We show how to (i) leverage Web data in order to discover and learn semantically meaningful attributes that are effective for re-identification and (ii) use this discovered attribute representation in conjunction with discriminatively trained matching techniques to obtain state of art

performance on a wide variety of re-identification datasets. We automatically construct a bottom-up attribute ontology, and learn an effective associated representation by large-scale mining of noisy but abundant content from social photo-sharing sites. Specifically, rather than asking an expert to define an ontology as in [103, 117, 153, 174] and the preceding Chapter 3, we discover an ontology automatically by clustering photograph meta-tags and social commentary. These clusters are used to train a large bank of detectors, resulting in a number of visually detectable attributes. Explicitly, this is in contrast to expert defined ontologies, which while intuitive to experts, may correspond to properties not possible to detect reliably with current vision techniques. This process is significantly more scalable than manually annotating data per surveillance site for attribute learning. Moreover, the greater volume and diversity of data used to train these automatically discovered attributes results in a more reliable and generalisable attribute representation than conventional attribute representation approaches on surveillance datasets can normally achieve. We validate our contribution by using our representation to evaluate our results on a set of four of the most challenging re-identification datasets.

Inspired by the success of attribute representations in other computer vision tasks, a recent line of work [103, 117, 153, 174, 106] has studied applying attributes to learn an informative representation for re-identification. The strategy has typically been to annotate binary or categorical clothing, object and soft-biometric properties on the training portion of a dataset, and then train models (such as topic models [117], SVM [103], or latent-SVMs [106]) to predict these mid-level properties based on some base low-level feature. Interestingly – *assuming attributes are reliably detectable* – only about twenty binary attributes are necessary to achieve unprecedented near perfect matching accuracy on challenging benchmarks [104]. The main bottleneck is actually one or both of robustness and accuracy with regard to attribute detection. This is hard to achieve because surveillance video is often of poor quality. However more fundamentally, it is challenging because obtaining sufficient annotated data to train reliable attribute detectors for each camera is prohibitively costly or impossible. In this chapter, we thus take a different approach to the attribute strategy, by mining attributes and attribute training data from social photo sharing sites. Automatically generating attribute detectors that both do not require manual annotation and are trained from sufficiently large scale data could be more scalable and generalisable. However, the challenge then becomes how to learn meaningful bottom-up attributes from large scale Internet data, given that such mining delivers highly noisy images and annotations.

4.2 Attribute Discovery

As ever growing amounts of visual data are being shared on the public Web, the computer vision community has begun to exploit this resource for obtaining large scale datasets and text or visual data mining [42]. Meanwhile, the availability of cheap crowd-sourced annotation has begun to make annotation of large-scale datasets more feasible [42]. However, crowd-sourced annotation at scale still incurs expenses in terms of time and human effort, and the results are often prone to bias and noise [163]. An alternative is to develop algorithms to mine data on the Internet [18, 106] with little or no human intervention. This may take the form of obtaining (noisily labelled) training data by image search using keyword query [30], or mining socially shared photos and associated tags/annotations [42].

With regards to attributes specifically, Chen *et al*.'s "Never-ending Image Learner", NEIL, [30] performed semi-supervised learning of attribute detectors based on large scale Internet image sets, starting with a small seed amount of annotated data. Meanwhile in the context of retail photos, [18] has clustered product photo annotations to automatically discover an ontology of putative attributes, for which detectors are then trained.

We employ a similar strategy to Berg *et al*. in [18], but we must discover attributes from deeply noisy and unconstrained data; rather than metadata and images from a noisy, but otherwise hand-curated website.

4.3 Discovering and Learning Attributes for Re-identification

In this section we outline how we first acquire a space of attributes from uncurated Internet data (Figures 4.2 on page 112 and 4.3 on page 113, Sections 4.3.1, 4.3.2), then how to train detectors for each attribute (Section 4.3.3) and fuse them with compatible representations for re-identification (Section 4.3.4). We include a schematic overview of our entire pipeline in Figure 4.1 on the facing page.

In order to alleviate the burden of annotating vast amounts of attribute training data, we first aim to acquire a large volume of *uncurated* and weakly labelled data from the Internet. Clearly, the kinds of photographs we might find online without a directed search stand a low probability of being immediately suitable for our purposes - Berg *et al*. [18], neatly summarise our problem as "identifying wheat from amidst a great deal of chaff".

Therefore we define a "broad" search query that is likely to return photographs that contain

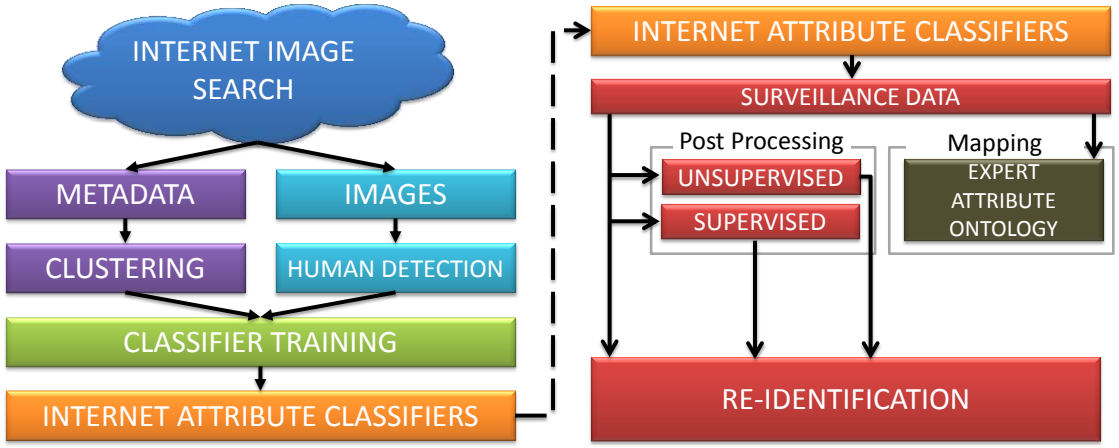


Figure 4.1: Schematic overview of our pipeline; Post-Processing modules such as distance-metric learning or domain-adaptation can be applied depending on the level of supervision available in order to boost "rank 1" or overall system performance as needed

.

depictions of people in everyday attire. We construct a boolean search query comprising of frequent synonyms of the word "person", such as "man", "woman", "pedestrian", etc, and combine this with multiple negative terms such as "car", "tree", "cat", and download 220,000 images with their associated metadata. This approach differs from most work on conventional recognition [171] where images are categorically and strongly annotated - or derived from a heavily curated source such as an eCommerce website selling a catalogued array of products. In our case, there is no guarantee that a photograph and associated metadata will have any meaningful semantic link, let alone whether or not the metadata refers to what we're really interested in: tags and keywords that describe the appearance of the people, if any, in the photograph.

The metadata for each photograph comprises of a variety of noisy but potentially useful information; location information is not used in our work, but present in approximately 8% of the photographs at least country-level, which could potentially be used to learn region-specific attributes in later work. For our purposes, we merge the photograph title and meta-tags, and employ common pre-processing measures to standardise the meta-text string somewhat; we tokenise and remove stop-words, remove numerical characters, and stem words to conflate semantically identical words to their common root. We do not apply a spelling-check so as to preserve any popular Internet vernacular, names or other bespoke allegorical terms that may be relevant or insightful at a semantic level in themselves, but may not have entered official spelling dictionaries. For example, a user's specific choice of tag for a city from all available toponyms may

reveal some information about the rationale behind the annotation; tagged images may also be somehow visually distinct as a result. Each photograph's meta-text is represented as a bag-of-words (BOW) histogram of bigrams with term frequency-inverse document frequency weighting (tf-idf) which ensures that salient words are more prominently represented.

Lastly, we constrain the constituent tokens of each bigram to being at least 3 characters long.

4.3.1 Discriminative text features from meta-text

As a first step to discovering latent attributes from the Internet data, we construct a BOW metatext representation with a vocabulary of $\approx 5,000$ unique bigrams (see Figures 4.2 on page 112 and 4.3 on page 113 for examples).

We construct an initial document-term matrix D, size $m \approx 69,000 \times n \approx 5,000$, where the ith row is an n-length vector d_i whose jth entry denotes how frequently a gram $gram_j \in G$ appears in metatext "document" $meta_i \in M$ obtained from each person detection $p_i \in P$. Each row d_i of D therefore represents a bag of words referring to a person detection and corresponding metatext; the jth element, representing individual gram counts for grams such as "blue" or "blue shirt".

This representational model is basic in that it assumes uniformity of importance across all terms and "documents" that introduces additional noise. In order to better emphasise grams that are potentially more meaningful than others, we apply the term frequency-inverse document frequency statistic (tf-idf) to D. We calculate the statistic for all entries in D as in the classic tf-idf formula in Eqs. (4.1, 4.2, and 4.3):

$$\mathrm{tf}(gram_j, d_i) = 0.5 + \frac{0.5 \times \mathrm{freq}(gram_j, d_i)}{\max\{\mathrm{freq}(w, d_i) : w \in D\}} \tag{4.1}$$

where w represents the maximum raw frequency of all terms in D.

$$\mathrm{idf}(gram_j, M) = \log \frac{N}{1 + |\{d \in D : gram \in m\}|} \tag{4.2}$$

The inverse document frequency function $\mathrm{idf}(t, M)$ is given in Equation (4.2), where N is the total number of person detections. The tf-idf is then finally constructed as:

$$\hat{\mathrm{D}}(t, m, M) = \mathrm{tf}(t, m) \times \mathrm{idf}(t, M) \tag{4.3}$$

$$\mathrm{S}(i, j) = |\hat{\mathrm{D}}(:, i) - \hat{\mathrm{D}}(:, j)|_2 \tag{4.4}$$

Where the operator : denotes columnwise selection in Equation (4.4 on the preceding page). We calculate the *L2* similarity matrix S as in Equation (4.4 on the facing page), between the frequency of the unigrams and bigrams, rather than using the Levenshtein distance on the second gram within each gram – this is unlike Marchesotti *et al.*'s approach in [126]. Next, we apply self-tuning Spectral Clustering [185, 130] to matrix S and select $N_a = 200$ clusters. Our intuition is that in our case it is the co-occurrence of the grams that is semantically relevant, not the similarity to other grams as represented by Levenshtein distance.

4.3.2 Person Detection

Many retrieved images are unsuitable for learning attribute-models suitable for surveillance because they contain landscape or objects instead of persons; or because persons are present but too close-up. To filter the data to obtain suitable images, we select Dollar *et al.*'s person detector [43][1] and employ both pre-trained models supplied by the authors to extract bounding boxes of people from this extremely varied collection of photos. This person detector is a vital component in dealing with the vast amount of noise inherent in the Internet-sourced images, since it affords us the ability to (i) determine if people are in an image with a measure of confidence, and (ii) be selective about how confident the detections we use for classifier training – in order to trade off data volume and label noise. After conservatively thresholding the person detection confidence, we are left with 69,000 person crops with corresponding meta-text features.

4.3.3 Classifier Training

Due to memory limitations related to kernel size, using traditional Support Vector Machines strategies for training large quantities of attribute detectors [104] was not tractable and therefore we select Linear Discriminant Analysis (LDA). Despite being a mature approach LDA still out-performs some contemporary machine learning methods, particularly for cases where there are many classes and comparatively few positive examples per-class. This, combined with being computationally less expensive and less sensitive to class imbalance, make it useful for our purposes. Using all 69,000 crops, we train an independent LDA model for each of the $N_a = 200$ discovered attributes. Finally we build a representation for any person's image X in an Internet-attribute semantic-space by stacking the positive-class posteriors from each detector into a N_a

[1] http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html

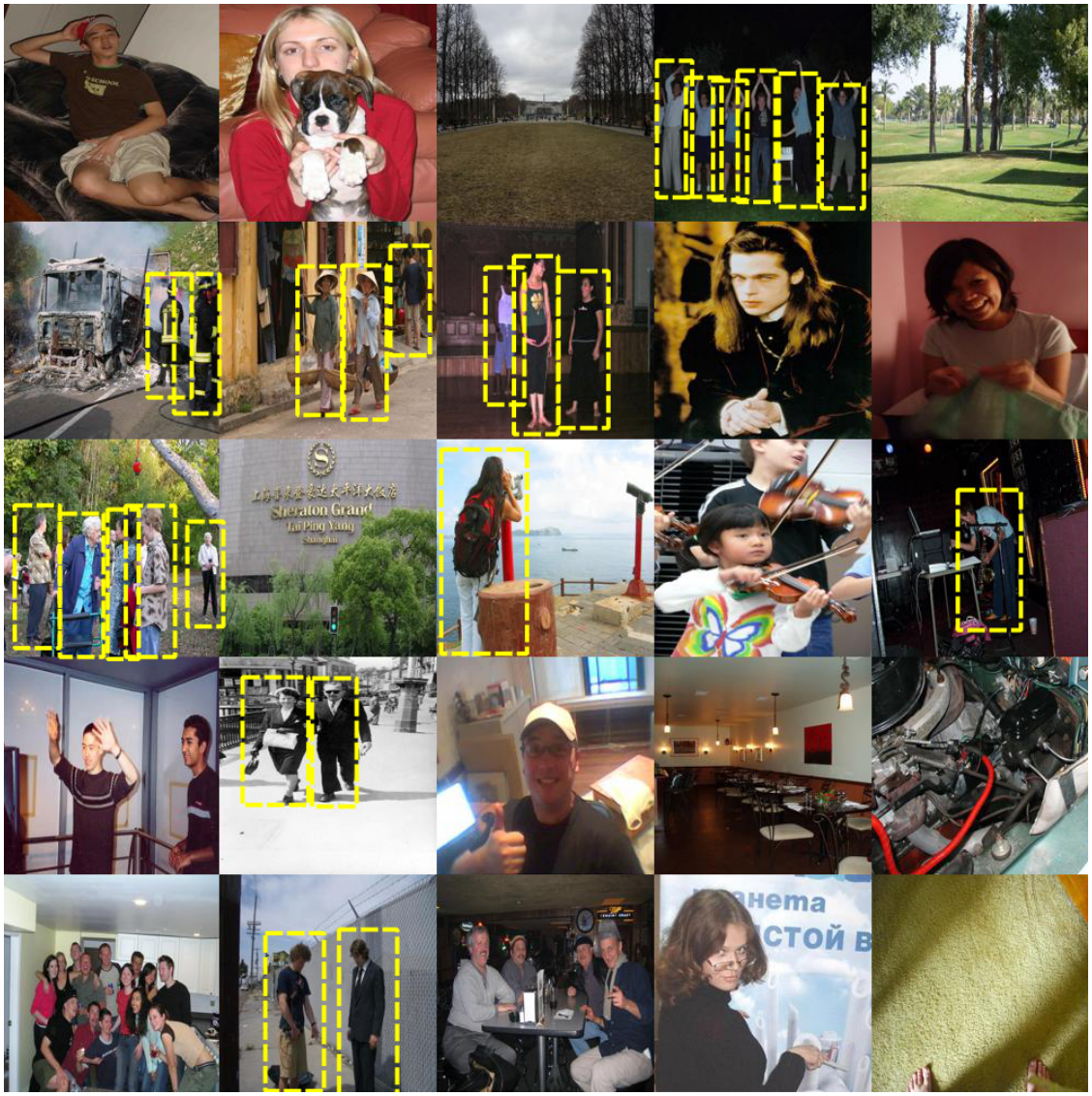


Figure 4.2: Uncurated images from our Internet search. Many images are unsuitable for surveillance attribute learning as they contain no people or are too close-up. For this work, we specifically filter out such images by discarding those photographs in which no full-body people can be detected reliably. As a qualitative illustration, even discounting people under occlusion in the above collage, there are potentially 25 candidates we might expect a person detector to locate in the 25 photographs pictured (yellow bounding boxes).

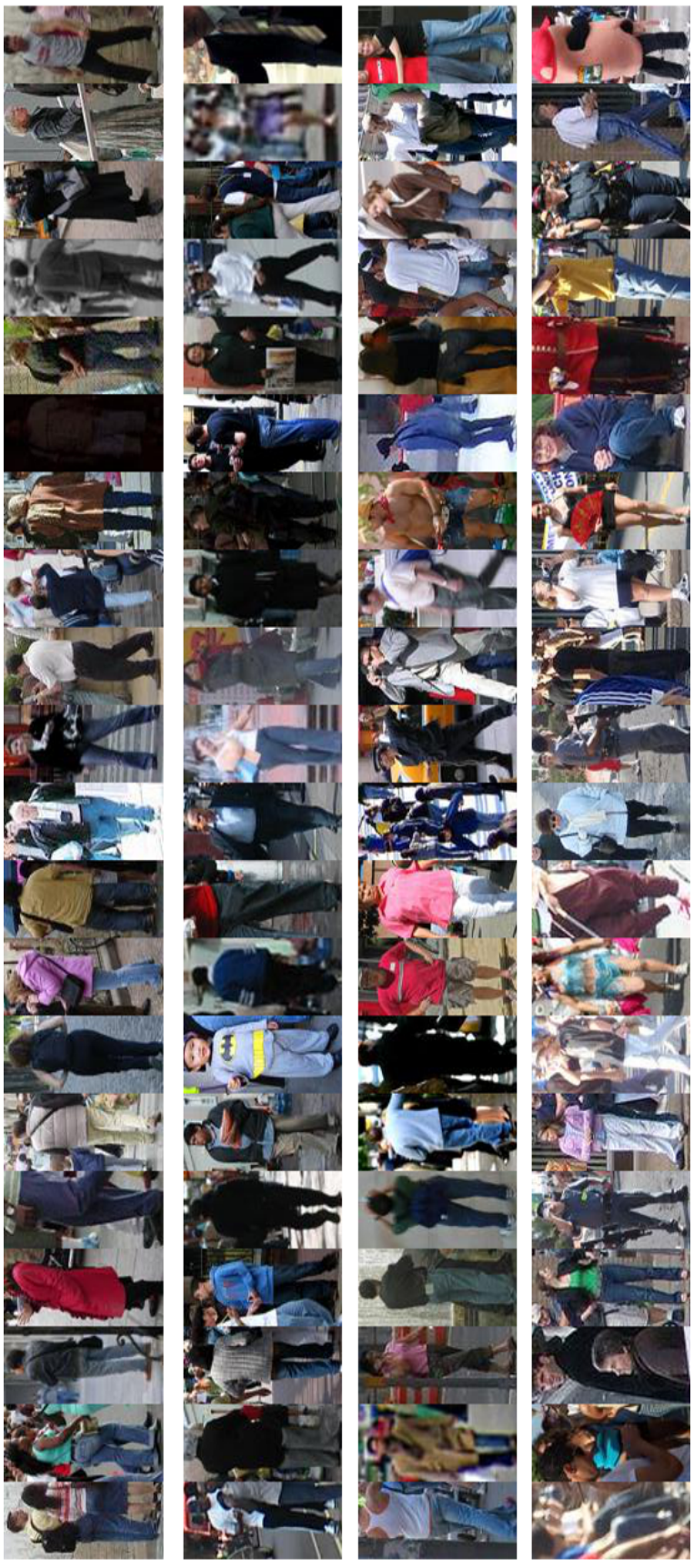


Figure 4.3: Person detections automatically extracted from *uncurated* Internet photographs. Each row comprises images from a discovered cluster, explicitly labelled on the left side. Noteworthy are the high variations in pose and appearance, fashion and background, as well as lighting and how the fashion varies according to location; for example despite the apparent seasonal variations within each cluster, it is noticeable that "New York People" have different tastes in apparel than "Paris People".

dimensional vector of Internet Attributes: $IA(X)$.

We train on Internet-sourced data, which one expects to have somewhat different statistics to typical surveillance crops. For example, surveillance crops typically come from lower quality cameras with more motion blur and compression artefacts. This may negatively affect the ability of our Internet data trained representation to effectively encode surveillance detections in practice. We therefore investigate applying unsupervised domain-adaptation to better align the Internet training data and surveillance test data. In particular, we align the projected subspaces of the two datasets, using Gong *et al.*'s geodesic flow kernel domain adaptation (DA) method [64].

4.3.4 Re-identification, Calibration and Fusion

The attributes obtained thus far are trained directly from discovered text clusters. There is variability in their reliability of detection based on image data, or their usefulness for re-identification. We therefore address learning a linear weighting $\mathbf{w}$ to rescale the attributes IA such that they are weighted according to their maximum utility for re-identification. Standard choices of optimisation criteria for re-identification include the first rank (R1) percentage, which reflects how often the first result in a ranked list is a perfect match to the probe, or expected rank (ER) or normalised area under curve (nAUC) of the cumulative match characteristic curve (CMC). We wish to enforce both a strong early-rank score, and good overall performance. To achieve this, we maximise the *product* of the CMC curve values $\hat{p}(k)$ at each rank k

$$\hat{P}_{\mathbf{w}}(k) = CMC_{\mathbf{w}}(k) = \frac{1}{n}\sum_{p=1}^{n}\mathbf{1}\left(k_p \leq k\right) \tag{4.5}$$

where k_p is the distribution of the ranks based on NN re-identification using $L1$ distances $D(IA_p, IA_g)$ between each attribute encoded probe $IA_p \in \mathcal{P}$ and all gallery members, $IA_g \in \mathcal{G}, g = 1, ..., n$. We denote an indicator function $\mathbf{1}$ that returns 1 or 0 following the evaluation of the parameters. Specifically we next use greedy search to select the weight $\mathbf{w}$ that maximises the following metric when used to scale each dimension/attribute a:

$$\max_{\mathbf{w}} \prod_{k=1}^{n} \hat{P}_{\mathbf{w}}(k) \tag{4.6}$$

Fusion with Low-Level Features Finally, we integrate our representation with metrics based on other low-level features. Specifically, we fuse BR-SVM [6] (trained on ELF features), SDALF [47] and our weighted Internet attributes after further discriminative training using KISS [94].

The resulting pseudo-metric's fusion weightings $\beta_{dataset}$ can be trivially selected with standard optimisation methods:

$$D(X_p, X_g) = d_{KISS}(IA(X_p), IA(X_g)) \tag{4.7}$$

$$+ \beta_{SDALF} \cdot d_{SDALF}(X_p, X_g) \tag{4.8}$$

$$+ \beta_{BRELF} \cdot d_{BRELF}(X_p, X_g). \tag{4.9}$$

For re-identification, we perform standard NN re-identification based on the fused metric in Eq (4.7), which we denote FUSIA, for FUSed Internet Attributes.

4.4 Experiments

We validate our contributions on four challenging public datasets, quantifying re-identification performance in the standard way [68] with CMC curve visualisations (CMCs), and expected-rank scores (ERs). CMC curves indicate the likelihood of a probe's true match appearing by the k^{th} rank, whilst ER represents the average rank of the true match to each probe – corresponding to the relevant metric of how far a human operator would have to search down a ranked list of matches before verifying the true target. High CMC values and ERs indicate better overall system performance.

4.4.1 Datasets

We tested the model using four publicly available re-id datasets: VIPeR [157], PRID [75], the QMUL underGround Re-IDentification dataset (GRID) [120] and CUHK [109], which provide 316, 200, 250, and 971 matched pairs respectively. These datasets cover a diverse variety of image sizes (in the region of [128x48] to [128x64].), typical view angles and camera conditions. For supervised learning experiments, we take a standard 2-fold partition approach to training and testing.

4.4.2 Person Detection, Representation and Domain Adaptation

We discard detections with confidence $c < 0.5$, in order to minimise false positives which degrade classifier performance. Cropped person detections are normalised to 128x48 pixels prior to feature extraction. For our visual features we employ the commonly used ensemble of local features[68] (ELF), which encodes both color and texture in 6 horizontal strips [147] for final features with 2784 dimensions, and reduce dimensionality to 100 with PCA; for feature fusion,

we also use symmetry-driven accumulation of local features (SDALF) as detailed in [47]. Note that SDALF provides a distance matrix directly, rather than a feature. For Domain Adaptation we have only one parameter to select, and use 10 dimensions as recommended by [64].

4.4.3 Visual Detectability of Internet Attributes

We first evaluate the visual detectability of the discovered Internet attributes. We train the binary attribute classifiers using semantic meta-text cluster assignments as labels, and randomly divide each cluster into training and validation partitions, containing 75% and 25% of the available data respectively. Across all folds and 200 attributes, average detection accuracy across the test-folds is 70.28%, which is significant considering that text-based attribute discovery is not guaranteed to produce attributes with visual correlates, and class imbalance between positive and negative classes may negatively impact discriminative learning models. Notably these numbers for detection reliability are comparable to 66-70% obtained using an expert-designed ontology purpose designed to be visually detectable and learned with extensively manual annotation of attribute training data [104].

4.4.4 Attributes as a Representation for Re-Identification

Figure 4.4 on page 119 summarises the re-identification performance of our complete algorithm, FUSIA, on all four datasets along with a variety of state of the art alternatives. The top plot shows CMC curves with our final model FUSIA - or Fused Internet Attributes, along with KISS [94], Binary-Relation SVM [6], SDALF [47] and saliency (eSDC) [186].

In the lower table we report scores obtained using our implementations of the cited methods in the first four rows. The remaining rows report results obtained from the cited works and blank results reflect where alternatives have not published results on a given dataset or format. In all cases we summarise with Rank 1 (perfect match rate), and expected rank. Our Rank 1 is comparable to state of the art alternatives, although not always best – however, our overall performance as evidenced by the CMC curves and their expected rank scores, outperform most alternatives by an often significant margin. This margin demonstrates the discriminative strength of our semantic attribute representation. Meanwhile the consistency of this margin across this wide batch of state of the art datasets demonstrates that the quantity and variety of source data is indeed leveraged to learn a highly generalisable representation.

Table 4.1 on the next page breaks down our method according to the different components and

contributions. Plain Internet attributes ("raw" IA) fail to outperform the ELF (upon which IAs are constructed). However, the full calibrated (weighted) and domain-adapted variant (IA), boosts overall re-identification performance dramatically to near state of art levels on VIPeR, GRID and CUHK, and maintaining comparable performance with other representations on PRID. Finally, applying metric learning to our attributes (IA-trained KISS) provides further improvement. The first three columns in Table 4.1 show the component metrics that are fused together to obtain the final result of FUSIA (final column).

	Component Scores			Comparison Scores			Final Result
Dataset	ELF [68]	IA (raw)	IA	KISS[94] (IA)	BRSVM[6] (ELF)	SDALF [47]	FUSIA
VIPeR	91.03	71.23	44.66	21.25	21.45	44.02	12.94
GRID	33.12	26.05	23.05	17.33	21.15	17.86	10.22
PRID	31.99	19.38	17.63	21.91	76.20	20.79	19.89
CUHK	161.39	138.41	128.13	72.25	43.28	72.96	38.09

Table 4.1: Breaking down re-identification performance by components of our full FUSIA model. See text of Section 4.4.3 for details. We report Expected Rank, lower scores are better.

4.4.5 Encoding Expert Attributes with Internet Attributes

A major advantage of our approach is that effectively, *unlimited numbers of person images can be obtained*. Thus, we would expect performance to improve with further application of computation time to crawling and learning more and better attributes. Nevertheless a disadvantage with our approach is that our attributes (Figures 4.2 on page 112 and 4.3 on page 113) are somewhat less easily interpreted by humans. Presented, for example, with the attribute "red shirt", the average human would be able to completely understand the concept regardless of the context being one of surveillance or shopping.

These conventional ontologies [104], by defining attributes such as "blue shirt" and "red shirt", map more clearly onto descriptive person search tasks whereas conversely, with our representation, attributes such as "Paris people" or "New York people" require more cognitive overhead to conceptualise and apply to the same task. To provide some insight into the mechanism of our contribution in this chapter, we illustrate the relation between these two interpretations of attributes: we use our framework to encode the VIPeR dataset in 200 dimensional IA representation, and then use existing VIPeR attribute annotation [104] to train a linear SVM mapping from a conventional attribute ontology to our representation. This corresponds to defining conven-

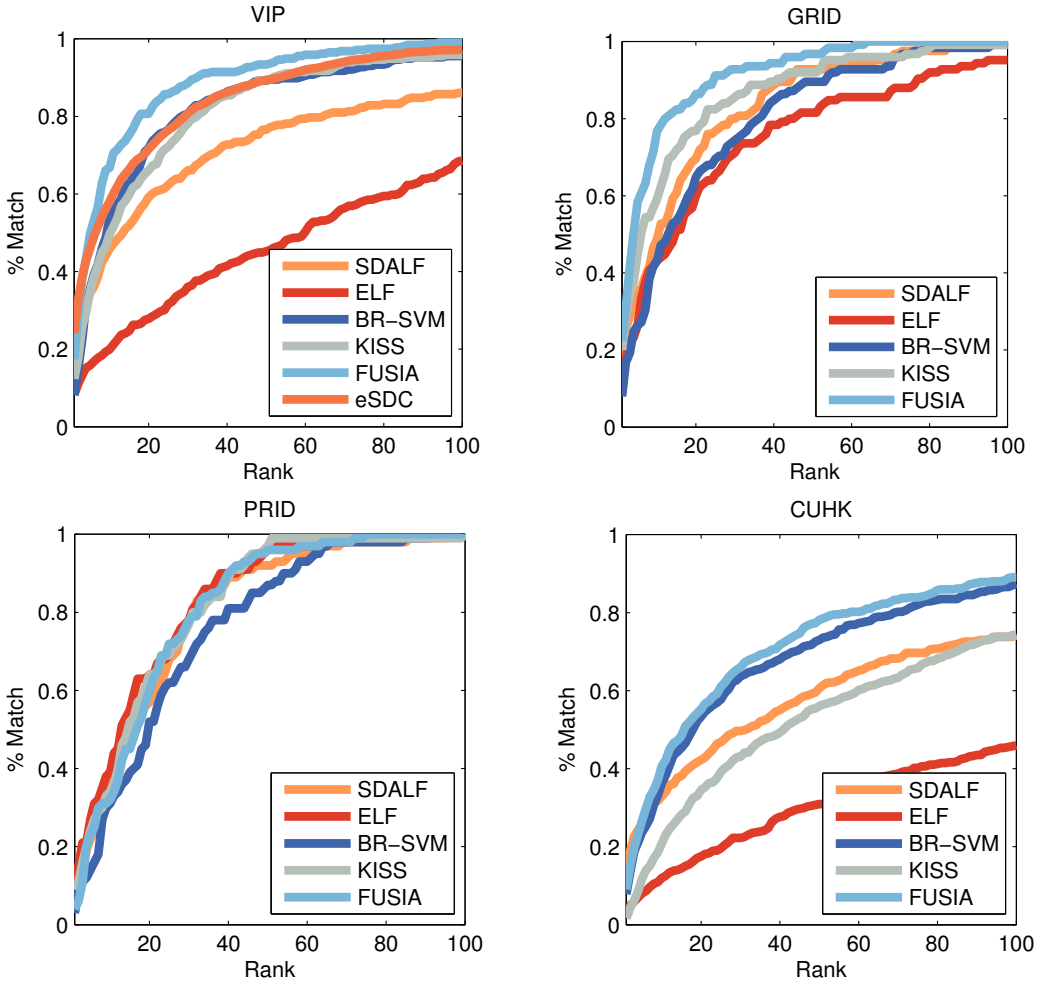
tional expert-defined attributes in terms of a linear combination of Internet attributes. Using "red shirt" and "blue shirt" as query terms, we demonstrate the top retrievals in our 69,000 person dataset in Figure 4.5 on page 120.

The results are compelling evidence that the approach we define in this chapter is able to encode information like "red shirt" or "blue shirt" within a somewhat higher-level attribute such as "Paris people". This has implications for applications beyond re-identification in surveillance: by connecting existing expert-defined attribute ontologies from surveillance to Internet data sources, we gain the ability to query Internet images for attributes without additional annotation of Internet data or training new classifiers for the Internet domain.

4.5 Discussion

We have shown in this chapter how effective mid-level semantic attributes can be discovered and trained from Internet data in an automated sense. These attributes are semantic by construction due to creation via mining of textual tags and comments, although they vary by how visually *obvious* they are in a human sense, they can be detected visually with comparable reliability to those attributes designed by human experts thanks to the practically unlimited quantity of Internet image data available for training. We demonstrate that this Internet attribute representation of person images is generalisable and discriminative for re-identification, a property that is unlocked through domain adaptation and metric learning, and furthermore is synergistic and amenable to fusion with conventional techniques.

However, a representation is only part of a full re-identification system and to realise full performance, a discriminatively trained matching model is required that requires pairwise annotation. In the following chapter we investigate one option for reducing the annotation cost for this step.



	VIPeR		GRID		PRID		CUHK	
Method	R1 ↑	ER ↓	R1 ↑	ER ↓	R1 ↑	ER ↓	R1 ↑	ER ↓
ELF [68]	0.08	84.27	0.13	26.42	0.11	18.26	0.04	160.55
BR-SVM [6] (ELF)	0.08	21.45	0.08	21.15	0.03	24.80	0.08	43.28
KISS [94] (IA)	0.12	21.51	0.20	14.73	0.09	19.27	0.02	73.16
FUSIA	0.17	13.39	0.22	9.55	0.04	19.90	0.09	38.59
SDALF [47]	0.16	44.02	0.16	17.86	0.03	20.79	0.12	72.96
eSDC [186]	0.24							
Liu *et al*. [112]	0.16							
RANKSVM [147]	0.15				0.10			
Hirzer *et al*. [76]	0.21		0.15					

Figure 4.4: Overall re-identification performance of our FUSIA representation versus alternatives, reported as CMC curves (top) and a table of Rank 1 and expected rank (ER) summaries (bottom).

Figure 4.5: Querying "red shirt" and "blue shirt" in our 69,000 non-labelled Internet-sourced person detections via a transfer mapping between our attributes and expert-ontologies from [104]

Chapter 5

Transferring Knowledge for Re-identification

In this chapter, we suggest that Internet Attributes from Chapter 4 are just one potential approach to the scalability problem; and whilst the intermediary representation afforded by our method in that chapter were proven to be a form of higher-level encoding of more interpretable attributes. We now move toward relaxing this strong assumption by investigating flexible multi-source transfer of re-identification models across camera pairs. Specifically, we show how to leverage prior re-identification models learned for a set of source view pairs (domains), and flexibly combine these to obtain good re-identification performance in a target view pair (domain) with greatly reduced training data requirements in the target domain.

Good progress can be made toward improving re-identification performance by using discriminative learning methods to directly learn a new representation as we demonstrated in Chapter 3. However, whilst this approach is promising it requires human curation in the form of expertly-defined labels for training, and also assumes that sufficiently diverse quantities of training data exist. In Chapter 4, we mitigate the data volume concern by introducing a way of mining semantically meaningful attributes from limitless supplies of Internet-sourced training images, as well as discover their compatibility and correlation with ontologies of expert-defined attributes such as those in Chapter 3.

Various recent approaches have made some progress in re-identification performance using discriminative learning techniques for both representation learning as detailed Chapter 3. However, these approaches are fundamentally limited by the requirement of extensive annotated training data for every pair of views. For practical re-identification, this is an unreasonable as-

sumption, as annotating extensive volumes of data for every pair of cameras to be re-identified may be impossible or prohibitively expensive.

5.1 Problem Definition

A central limitation of existing discriminative learning approaches, is that they are most suited to closed-world benchmark problems than realistic open-world scenarios. In particular they require many pairs of person images annotated by same/different, *for each camera pair* between which the system is required to operate. This is reasonable for training/testing splits on benchmark datasets that are already exhaustively annotated by person identity. However it is highly impractical for real-world use, where there may be very many pairs of cameras in a given network, *each* requiring exhaustive annotation – making this "calibration" requirement of such a system impossible or prohibitively expensive. Ideally, we would like to deploy a re-identification system between a pair of cameras with minimal calibration/training annotation. What a system learns from annotations of one camera pair should be exploited by another pair without requiring exhaustive annotation in the new pair.

This is an issue in *transfer learning* [140, 45, 83]. Transfer learning has been used to good effect in numerous classical computer vision problems, for example object categorisation [83, 151]. The motivation is typically to scale systems to many classes [83] or domains [151, 45] without requiring prohibitive amounts of training data. While transfer learning is already an important issue in classical vision tasks, it will turn out to be even more central to the re-identification problem. This is because since *pairs* define domains in this context, it is unreasonable to collect exhaustive training data for a quadratic number of domains.

Transfer learning is already important for many classical vision problems with multiple classes or domains. However it is critically important for re-identification because the number of domains (camera pairs) is *quadratic in the number of cameras*. Therefore obtaining exhaustive training data for each domain is even more impractical than for conventional vision applications, and transfer learning becomes critical. Nevertheless, no prior re-identification studies have addressed this issue, relying solely on benchmark datasets with sufficient annotated data in each camera-pair of interest.

In this chapter we relax the practically unrealistic assumption of exhaustive training data within each domain by generalising recent ideas in learning re-identification [6] and SVM trans-

fer learning [83]. Specifically, we consider re-identification based on binary-relation learning [6, 96], and show how to generalise this approach to achieve effective cross-domain learning by combining non-linear decision boundaries from source domains to create a more accurate target domain re-identification classifier. In this way we are able to improve on within-domain learning both for sparse and even non-sparse training data volumes. Moreover we show how to achieve this while systematically avoiding negative transfer, even when there are multiple and irrelevant source domains.

5.2 Transfer Learning for Re-identification

Learning approaches to re-identification typically learn distance metrics [77, 192, 94], or model-based matching procedures such as boosting [68] and ranking [147] based on annotated training pairs. These have recently improved state of the art re-identification performance significantly [77, 6]. Another line of research learns mid-level attributes [101] to replace or augment low level features. In this case inter-camera invariance is obtained via the generalisation performance of learned attribute classifiers. However, this only applies within domains where annotated attribute data are available. The recently proposed binary relation learning approach [6, 96] obtains state of the art re-identification results by exploiting strong SVM classifiers trained to make same/different judgements on pairs of images. This strategy does not assume that instances of the same person are more similar than instances of different people, and instead implicitly learns the mapping between appearance in pairs of training cameras.

A serious issue with all these approaches is that *they do not generalise well across domains* (different re-identification view pairs; see Section 5.3.5); and hence require extensive volumes of training data for *each pair* of cameras between which re-identified is to be performed. This is possible for benchmark scenarios, but unreasonable in practice.

5.2.1 On Cameras and Domains

In this work we consider a camera *pair* to make up a *domain*, and this should not be confused with some other studies which consider a particular *camera* to be a domain [151]. For classification [151] and detection [45], an individual camera encompasses the notion of a domain because a camera's parameters impart a systematic impact on the observations, which the model must learn to interpret. However in re-identification, a model's task is to infer something about pairs of

Figure 5.1: Examples from all of the datasets we use in our experiments, from the top: VIPeR, PRID, GRID, and CUHK. Note the dramatic appearance variations in both the people and backgrounds; as well as how image quality varies.

observations, and the systematic impact of the environment is therefore defined by the pair of cameras.

5.2.2 Transfer Learning

Only very recently has transfer learning for re-identification begun to be considered [109, 191]. However these studies consider only improving within-domain (camera pair) re-identification by transferring knowledge learned from one group of people to help identify another group of people. This is intrinsically a much more restricted scenario than the more general and useful case of transferring across domains to permit re-identification in a new camera pair with sparse annotations.

A central issue in transfer learning [140] is that of *from where to transfer*. When there is only one source of information available, and that source is known to be highly relevant to the task of interest, then transfer learning is much simpler than in the more general and realistic case where there are multiple sources of information of greatly varying relevance. In this latter case, it is non-trivial to design models which avoid negative transfer [140]. Our problem of transferring mappings across camera pairs falls squarely into the latter more difficult case. Since the relevance of one camera pair to another depends on similarity in their viewing angles and lighting, many pairs will not be similar and working out from where to transfer is of critical importance.

5.2.3 Negative Instance Selection

In our framework and many other methods [147, 188] which are trained on pairwise data, there is the issue of which examples to choose among the quadratic number of negative instances. The work of [6] presented an analysis showing diminishing returns but increasing computational cost beyond 10 negative per positive instances. However, choosing instances randomly means that most negative pairs will be far from the decision boundary and convey no extra information (see Figure 5.2 on page 127). This means that: (i) computation is wasted, (ii) performance is suboptimal because many informative negative pairs will be missed, and (iii) this is not scalable in terms of human annotation.

5.2.4 The Approach

We address all the mentioned issues by generalising the state of the art binary relation approach to re-identification [6], but tackle the new challenges in addressing the training data requirements

via multi-source transfer. There are many potential approaches to transfer learning [140], but in this study we will develop a SVM multi-kernel learning (MKL) [46, 83] transfer strategy. This will allow us to integrate multiple source domains of unknown relevance, while avoiding negative transfer via an inter-kernel sparsity regulariser

We make the following specific contributions: (i) Framing the problem of generalising re-identification as a domain-transfer problem; (ii) Developing a specific framework for domain-transfer re-identification for multiple domains of varying relevance by way of expressing the task as a SVM multi-kernel learning problem; (iii) Revealing the limitations of existing approaches to re-identification by way of a systematic and quantitative cross-domain evaluation; and (iv) Extensive evaluation of our proposed method on four of the largest public re-identification datasets available.

5.2.5 Concept Illustration

To provide intuition before introducing the details of the proposed method, Figure 5.2 on the next page provides an schematic illustration of our re-identification transfer learning framework. In this illustration, the feature space within each camera is one dimensional. A domain, consisting of pairs of observations made by two cameras, can thus be represented as a point on a two dimensional plane. Pairs of cross-view images corresponding to the same person are shown with circles, and pairs corresponding to different people with crosses. Binary-relation [6] based re-identification is the strategy of learning a decision boundary in this space (Figure 5.2 on the facing page, blue lines). In an easy re-identification scenario, the feature-space is the same in each view, so distinguishing true pairs from false pairs requires only a simple decision boundary (Figure 5.2 on the next page(a)). In a realistic scenario, there will be a non-trivial and unknown transformation [146] in feature space from one camera view relative to another (Figure 5.2 on the facing page(b) and (c)). In this case a strong non-linear classifier could learn the decision boundary separating true from false pairs, and hence an implicit inter-camera mapping.

In this illustration, we assume there are three source domains (camera pairs; Figure 5.2 on the next page(a)-(c)) for which annotated data (red and green symbols) is plentiful, and good binary relation based re-identification models have been learned (blue lines). Now suppose we wish to deploy our re-identification system to a new location where we can only annotate a very limited amount of training data. With limited data, a re-identification classifier learned in the conventional way – solely from local data – will be much less accurate, clearly misclassifying

many regions of the input space (Figure 5.2(e), unlabeled grey symbols on the wrong side of the decision boundary). In contrast, a re-identification classifier taking advantage of our domain transfer framework will realise that the limited data is best explainable by the model learned from the second source domain (Figure 5.2(b)), and borrow that classifier's strength to help learn a much more informative and accurate boundary than is possible using local data alone (Figure 5.2(d) vs (e)). (The intuition for how this works is finding a source domain classifier or combination thereof which fit the few available data points in the target domain). Finally, note that simple averaging of all the source classifiers is insufficient: in this example the mean of source classifiers (a)-(c) is very similar to classifier (a) which will be wrong for the target domain (d). We shall validate these intuitive observations experimentally in our experiments (Section 5.3.4).

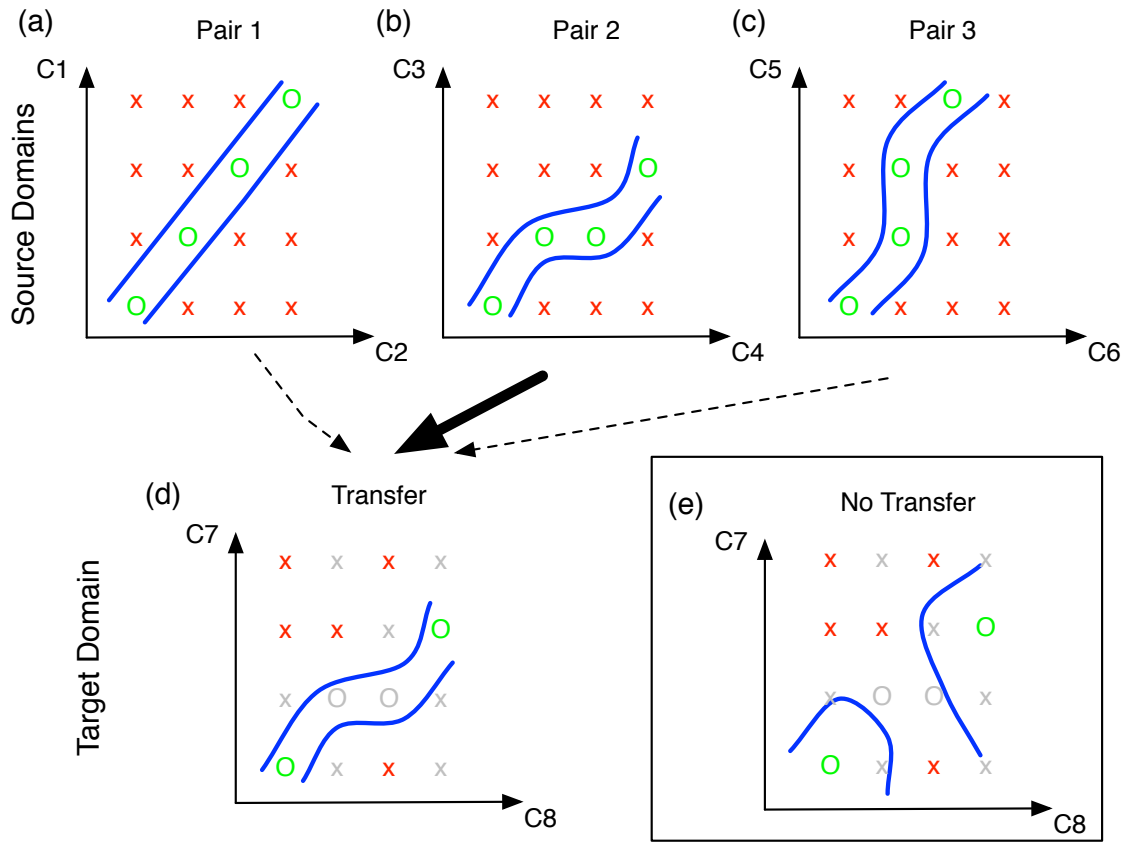


Figure 5.2: An illustration of how domain transfer can assist re-identification. We simplistically represent domains as a pair of one-dimensional axes, where each axis C represents a different camera. Symbols (O, X) indicate same/different pairs, grey symbols are un-annotated data points and blue lines indicate decision boundaries. Here, auxiliary domains (a-c) each provide useful information (arrow weighting) to the target domain classifier (d), in conjunction with some annotation for calibration. Our method avoids the failure case (e) where an erroneous decision boundary is formed.

5.2.6 Within Domain Re-identification

Training

We first consider the case of learning to re-identify people within one particular domain corresponding to a camera pair a and b. Here we largely follow a binary-relation learning approach [6, 96], but review the method for completeness. We assume training data $\{\mathbf{x}_i^a, z_i^a\}_{i=1}^{N_a}$ describing N_A people observed in camera a, and $\{\mathbf{x}_j^a, z_j^b\}_{j=1}^{N_b}$ describing N_B people appearing in camera B, where $\mathbf{x}$ represents a feature vector, and z indicates the identity of each person. From this data we can generate:

- A set of cross-camera positive pairs of the same person:
  $\{y_k = 1, \mathbf{x}_k = [\mathbf{x}_i^a || \mathbf{x}_j^b]_k\}, \forall (z_i = z_j)$,
- A set of cross-camera negative pairs of different people:
  $\{y_k = -1, \mathbf{x}_k = [\mathbf{x}_i^a || \mathbf{x}_j^b]_k\}, \forall (z_i \neq z_j)$,

where $[\cdot||\cdot]$ denotes concatenation and $k = 1 \ldots N$ indexes observation pairs $\mathbf{x}_k$. Note that there are a quadratic number of negative pairings, and actually constructing all pairs is typically prohibitive, so using a random subset of negative examples is typically adopted [6, 147].

Specifically, to sample negative instances we take each positive instance $i \in A$ from camera A in turn and at random uniformly sample, (without replacement), 10 negative instances $j \in B$ from camera B with the constraint $j \neq i$. To learn a re-identification model, we train a classifier on pair data $\{y_k, \mathbf{x}_k\}_{k=1}^{N}$ to distinguish matching pairs from non-matching pairs. This can be formalised as a support vector machine learning problem as:

$$\begin{aligned} &\min_{\mathbf{w},\xi} \|\mathbf{w}\|^2 + \frac{C}{N}\sum_{k=1}^{N} \xi_k, \\ &\text{s.t. } y_k \mathbf{w}^T \phi(\mathbf{x}_k) \geq 1 - \xi_k, \ \forall_k, \end{aligned} \tag{5.1}$$

where C parametrises margin penalty, $\phi(\cdot)$ is a non-linear mapping, and we maximise the margin subject to the soft constraint (non-negative slack variable ξ_k) that true pairs should be positive and false pairs should be negative.

Discussion

Note that this objective (Equation 5.1) pursues positive true pairs and negative false pairs, without any assumption of their visual similarity/dissimilarity. With the RBF kernel, binary-relation SVM implicitly learns an arbitrarily complex transformation mapping between cameras, e.g.,

uncovering their lighting [146] or view transformation, as well as relative relevance for each feature within that domain. In contrast, the common RankSVM [147] approach has two limitations: (i) it only models a first-order weighting of features, without considering their covariance, and (ii) it operates under the explicit assumption that true pairs should be more similar than false pairs (i.e., Figure 5.2 on page 127(a)). In practice this means that for camera pairs which deviate sharply from a simple linear transformation model (e.g., Figure 5.2 on page 127(a)) to a more complex transformation (e.g., Figure 5.2 on page 127(b) or (c)), binary relation SVM outperforms RankSVM, as shown in [6]. Mahalanobis metric learning objectives [77, 94, 192] are more powerful than RankSVM in modelling feature covariance, however they also still assume that true pairs are more similar than false pairs.

On Transferred Re-identification

For online re-identification of persons across cameras, putative pairs of images are concatenated $\mathbf{x}_* = [\mathbf{x}_*^a, \mathbf{x}_*^b]$ and the score of a test pair $\mathbf{x}_*$ is is evaluated as $f(\mathbf{x}_*) = \mathbf{w}^T \phi(\mathbf{x}_*)$. The pair can be classified as same or different via sign$f(\mathbf{x}_*)$, or the continuous score itself can be used to relatively rank putative matches. Given this re-identification framework, we next address how to transfer learned models across domains.

5.2.7 Domain Transfer Re-identification (DTR)

Training

Assume a set of source domains $s = 1 \ldots S$ are given, for which we have learned re-identification models as per Section 5.2.6. To leverage the learned experience of these domains in a new target domain t, we take the strategy of multi-kernel learning [8]. Each source domain s can be seen as providing a score $f_s(\mathbf{x})$ indicating its confidence that a given pair $\mathbf{x}$ is a matching pair under the model of that domain. We therefore formalise a domain transfer prediction task, which classifies a pair $\mathbf{x}$ in the target domain, taking into account both target and source domain knowledge, as:

$$\begin{aligned} f_t(\mathbf{x}) &= \overline{\mathbf{w}}^T \overline{\phi}(\mathbf{x}), \\ &= \mathbf{w}_t^T \phi_t(\mathbf{x}) + \sum_{s=1}^{S} \mathbf{w}_s^T \phi_s(f_s(\mathbf{x})), \end{aligned} \tag{5.2}$$

where parameters $\overline{\mathbf{w}} = [\mathbf{w}_t, \mathbf{w}_s]$ to be determined weight the relative informativeness of the target domain and each source domain knowledge.

Given this task formulation, the within-domain learning objective in Eq (5.1 on page 128) can be generalised to the case of domain-transfer learning to estimate $\overline{\mathbf{w}}$ as:

$$\min_{\overline{\mathbf{w}}} \Omega(\overline{\mathbf{w}}) + \frac{C}{N}\sum_{k=1}^{N} L(\overline{\mathbf{w}}, \mathbf{x}_k, y_k) \tag{5.3}$$

where L denotes the hinge loss

$$L(\overline{\mathbf{w}}, \mathbf{x}, y) \quad = \quad \left|1 - y\overline{\mathbf{w}}^T \overline{\phi}(x)\right|_+ \tag{5.4}$$

and $\Omega(\overline{\mathbf{w}})$ denotes the weight regulariser. Note that [83] use a linear kernel ϕ for computational tractability. In our case, because the problem is binary unlike in [83], we are able to use the RBF kernel instead without great computational penalty. This is indeed necessary because we need to learn a complex transformation.

Evaluating Domain Relevance

An important issue for domain transfer in the general unconstrained case is that we do not know in advance which source domain is going to be relevant, and indeed the majority are likely to be irrelevant. For this reason we seek a sparse solution for the optimisation problem in Eq (5.3). Previously L_1 norm regularisers have been proposed to provide sparsity across kernels. However this is hard to optimise effectively [8]. The L_p $(1 < p \leq 2)$ norm regulariser has recently been shown to effectively induce sparsity while providing significantly easier optimisation [137]. We therefore take the $(2, p)$ group-norm as the regulariser: providing L_2 regularisation within domains, while encouraging L_p sparsity across the set of $S+1$ kernels which reflect the cues from the target domain and the S source domains:

$$\begin{aligned} \Omega(\overline{\mathbf{w}}) \quad &= \quad \frac{1}{2}\left\|\overline{\mathbf{w}}\right\|_{2,p}^2, \\ &= \quad \frac{1}{2}\left\|\left[\left\|\mathbf{w}_t\right\|_2, \left\|\mathbf{w}_1\right\|_2, \ldots, \left\|\mathbf{w}_S\right\|_2\right]\right\|_p^2. \end{aligned} \tag{5.5}$$

Explicitly, the $(2, p)$ group-norm applies an L_2 norm within kernels (sources), but L_1 across sources. In other words, the sparsity inducing L_1 regulariser will try to reduce an entire source to zero if possible, but the L_2 regulariser will not do the same to individual weights *within* a source. This is good for discounting irrelevance at the level of sources rather than individual features, avoiding negative transfer because any source kernels which mismatch the available target domain data will be allocated zero coefficients. Expressed in this form, we can exploit existing efficient stochastic gradient-descent algorithms [46] for solving the cross-domain re-identification learning problem in Eq (5.3).

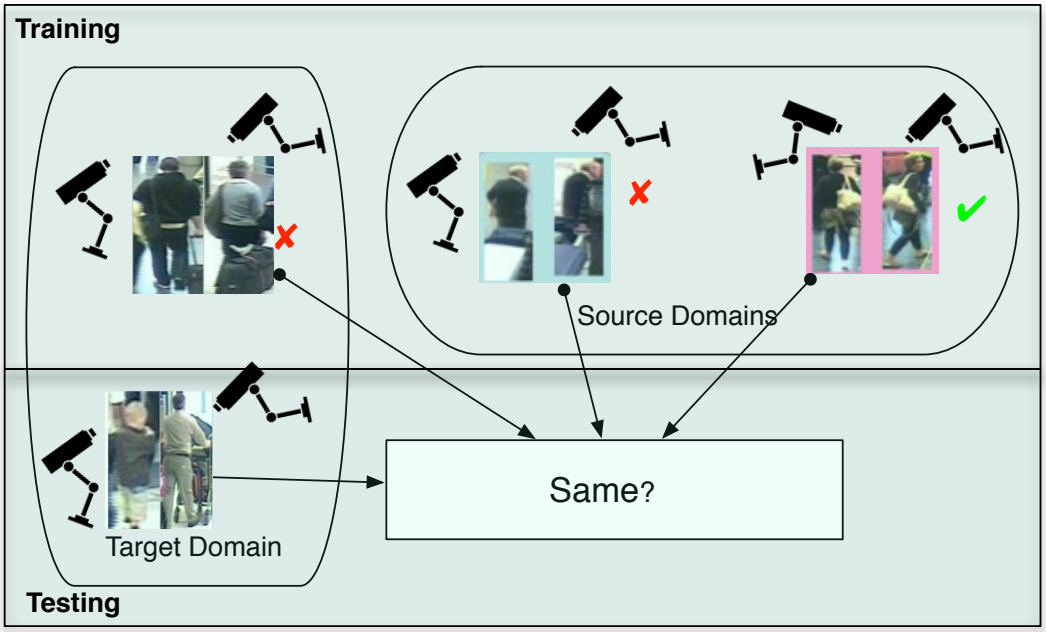


Figure 5.3: Schematic overview of our framework for transferring knowledge from previously trained camera pairs onto a new camera pair.

5.3 Experiments

5.3.1 Feature Extraction

The main imagery feature that we will use with our DTR model is the 150 dimensional HSV colour descriptor as detailed in [6]. Additionally we compared the commonly used ensemble of local features (ELF) which encodes both colour and texture in 2784 dimensions as detailed in [68, 147]; as well as symmetry driven accumulation of local features (SDALF) as detailed in [47]. Note that SDALF provides a distance matrix directly, rather than a feature encoding.

5.3.2 Experimental Settings

We tested the model using the four largest publicly available re-identification datasets: VIPER [68], PRID [75], GRID [120] and CUHK [109], which provide 316, 200, 250, and 971 matched pairs respectively. These datasets cover a diverse variety of image sizes (in the region of [128x48] to [128x64].), typical view angles and camera conditions (Figure 5.1 on page 124). We evaluated cross-domain re-identification performance on these datasets in four "leave one dataset out" folds. In each case we considered three datasets as source domains and the fourth dataset as the target domain. For the source domains we learned within-domain re-identification models with

all available data for each (Section 5.2.6). For the held out domain, we performed 2-fold cross-validation, training the domain transfer model on half (or less) of the data (Section 5.2.7), and using the held out half for testing. For testing, we consider the matched pairs between cameras within the domain, taking each person in turn (probe) and matching them against the people in the other camera (gallery). Within the source domains, SVM slack parameter C was cross-validated to optimise expected rank. In the target domain we set $C = 10$ throughout. We fixed the RBF kernel parameter γ to the median of each distance matrix. For the SVM methods we select 10 negative examples per positive pair.

5.3.3 Evaluation

As baselines we consider where relevant three non-learning methods and three learning methods. For non-learning methods we consider: (i) HSV features [6], (ii) ELF [68] and (iii) SDALF [47]; in each case with nearest neighbour (NN) matching and Euclidean distance where relevant. For learning methods, we consider:

ATTR: Re-identification using Euclidean NN matching on learned mid-level attributes [101] from ELF [68] features.

BR-SVM: Binary-relation based re-identification using SVMs [6, 96]. Note that BR-SVM has already been shown to decisively outperform the commonly applied RankSVM [147, 191] and prior metric learning methods [192].

DTR: Our proposed new Domain-Transfer re-identification model, using multi-kernel learning.

We evaluate re-identification performance using two metrics: For visualisation, the normalised Cumulative Matching Characteristic (nCMC) curve, which indicates the probability of the correct match to a probe image appearing in the top n results from the gallery for varying n[1]. For quantitative summarisation, we use the expected rank (ER) metric [68, 6], which is the mean rank of the true result[2]. This metric has the advantage that it reflects a physically meaningful quantity, which is how many items an operator has to scan in a ranked list before reaching the true match for the probe, and hence the average time it takes a human operator to find the true match using such a system [68].

[1] Here, higher curves are better; enclosing an area of 1 is perfect; and an area of 0.5 is random

[2] Lower is better; a mean rank of 1 is perfect; and a mean rank of half the gallery size is random

5.3.4 Domain Transfer Experiments

Domain transfer compensation for a lack of target domain data

We first evaluate re-identification performance as a function of target domain training data volume. Figure 5.4 on the next page summarises the expected rank (ER) of each model for logarithmically varying volumes of training data. Also shown (flat lines) are the performance of LLF models SDALF (red), HSV (blue) and ELF (black). Clearly performance for the learning models degrades with sparser training data (Figure 5.4 on the following page, ER of learned models higher to the right). However, our proposed DTR model (magenta) systematically outperforms the within-domain BR-SVM model [6] (green), especially with increasingly sparse data. We obtain between a margin of improvement over BR-SVM of 5-20%, 6-16% and 6-17% for VIPER, GRID and CUHK respectively. Meanwhile we obtain a margin of improvement over SDALF of up to 70%, 5%, 25% and 31% for VIPER, GRID and CUHK. At some point, for all learning models, the data will be sufficiently sparse that LLF approaches will be best. However DTR's margin over BR-SVM, means that standard LLFs can be outperformed with less training data than before. DTR model outperforms the best LLF with down to 1/16th data for VIPER, 1/4 data for GRID and 1/8th data for CUHK. Importantly, performance of DTR is usually dramatically better than simple nearest-neighbour on HSV (blue), which is the feature on which DTR was trained. Note that our weaker result on the PRID dataset can be understood by the generally poor performance of the HSV feature used by our DTR in this domain (see Section 5.3.5). This could in general be ameliorated by including other feature types within our MKL framework.

These results are also visualised in Figure 5.5 on page 135, showing the CMC curve for each domain and data sparsity condition (line-style), of BR-SVM-based re-identification versus our domain-transfer model (colour). The magenta CMC curves representing the transfer condition enclose the green non-transfer curves in each case. Finally, for GRID and CUHK we observe that even with the maximum volume of training data, transfer learning is still able to improve performance (Figure 5.5 on page 135, solid magenta CMC curves enclosing solid green CMC curves; Figure 5.4 on the next page, magenta curves under green curves).

Some visual examples of the improvement provided by our DTR approach over BR-SVM in each dataset are shown in Figure 5.7 on page 141. In each case, the correct match to the probe is highlighted in green and the upper rows show the ranked matches by DTR versus ranked matches by BR-SVM in lower rows. Finally, Table 5.3 on page 139 summarises some accuracies of each

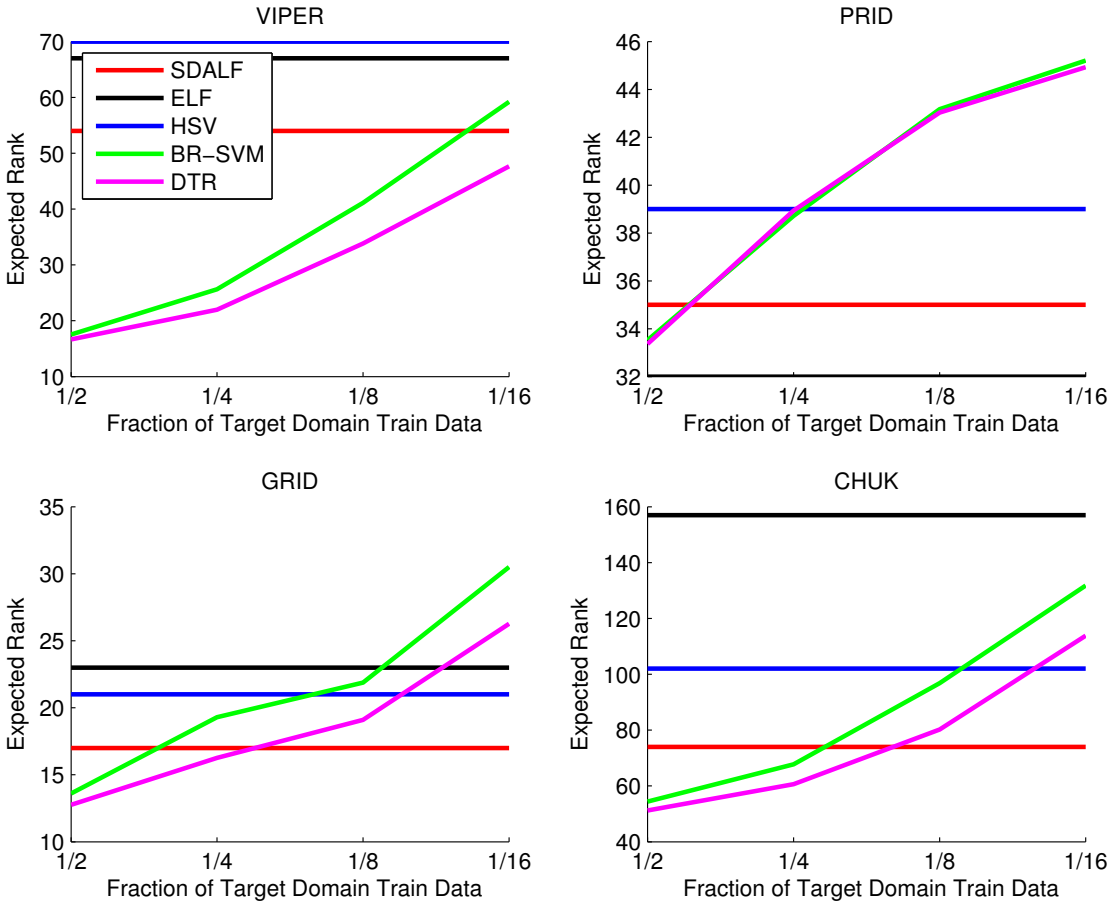


Figure 5.4: Re-identification performance as a function of volume of training data. Lower expected rank is better. Each dataset is evaluated as a leave-one-dataset out domain transfer problem. Our proposed DTR model systematically outperforms BR-SVM within-domain learning.

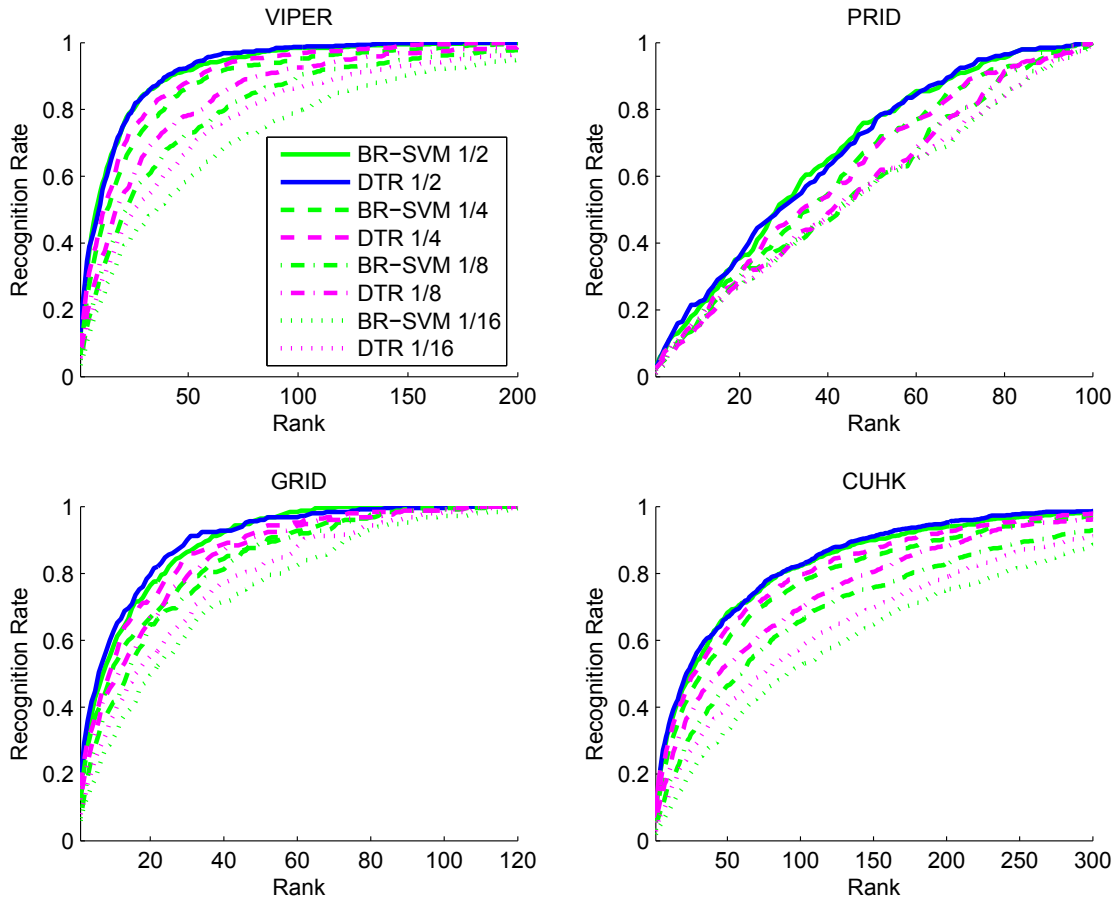


Figure 5.5: CMC curves for re-identification with and without transfer. Each line-type illustrates a different volume of training data. In each case the transfer CMC curve encloses the non-transfer curve.

---

method at different ranks under the various conditions. In the majority of cases DTR clearly outperforms BR-SVM.

Cross-Domain Analysis

To provide some insight into the cross-domain results above, we present some analysis of the affinity between the major re-identification datasets by way of the learned weights for each kernel. Figure 5.6 on the following page plots the weights for re-identification for each target domain (rows) against the data source (columns). As expected, each dataset is highly relevant to itself (strong diagonal). Cross-dataset transfer is illustrated by the off-diagonal weights. It is evident that the VIPER re-identifier is relevant to assist both GRID and CUHK, but not PRID. Interestingly, there is some degree of transferability between VIPER, GRID and CUHK. However, the PRID dataset is neither useful as a source for any others, nor making use of any others as a

source. This reflects the previous (Figure 5.4 on page 134) results showing that the transfer performance for PRID was no better than the local only performance. Nevertheless, it is reassuring that in this case of an irrelevant source, the sparsity prior of our transfer framework was able to apply zero weighting (Figure 5.6) and hence avoided automatically negative transfer (Figure 5.4 on page 134).

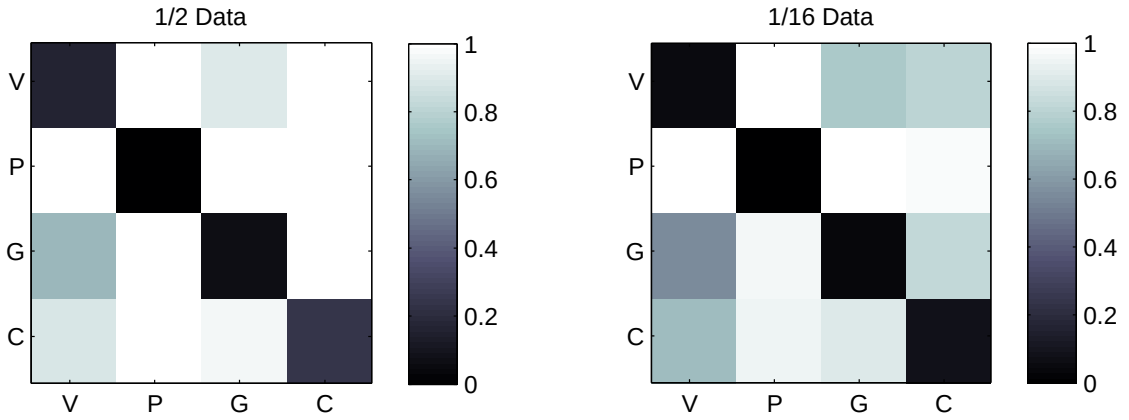


Figure 5.6: Cross-dataset affinity for re-identification. Darker blocks indicate a stronger cue.

5.3.5 Additional Analysis

We next provide some additional analysis about the existing models and datasets to provide some insight into the domain transfer problem, and further validate our contribution as illustrated in Sections 5.2.5 and 5.3.4.

Generalisation of low-level features

To investigate the generalisation of low-level features, we perform re-identification using non-learned nearest-neighbour matching on the four datasets. The results are shown in Table 5.1 on the facing page, expressed as expected rank. The best results are highlighted in bold, and the worst in red. The important point to note here is that the best and worst results using low-level features vary significantly on different domains. That is, the rankings obtained by different feature types are not uniformly good across domains. This highlights in turn that just making a single selection of "good" feature for re-identification and expecting similar performance on all domains is not plausible. Therefore, leveraging learning based methods to adapt to the appearance of a given camera view is critical. We note that while SDALF [47] is the most effective feature overall, it is extremely computationally extensive to extract and thus of limited suitability for practical real-time applications.

	HSV[6]	SDALF[47]	ELF[68]
VIPeR	70.24	**53.64**	67.73
PRID	38.91	34.85	**32.50**
GRID	20.64	**16.70**	23.18
CUHK1	101.72	**73.70**	156.86

Table 5.1: Low-Level Features (LLFs) often do not generalise across domains. Columns are LLFs used in NN re-identification on four public datasets (rows). We report Expected Rank (ER), lower scores are better. Bold scores are best; red scores are worst.

BR-SVM[6]	VIPeR	PRID	GRID	CUHK
VIPeR	**16.17**	50.23	39.01	166.11
PRID	155.23	**34.35**	59.70	240.72
GRID	119.38	49.17	**11.60**	202.55
CUHK	96.51	48.93	47.39	**52.24**

ATTR[101]	VIPeR	PRID	GRID	CUHK
VIPeR	**48.19**	43.38	26.22	185.61
PRID	98.82	**26.06**	39.01	201.50
GRID	94.28	46.69	**21.82**	194.29

Table 5.2: Learning-based re-identification methods may transfer "blind" and retain some utility on untrained datasets but performance is penalised. Rows are training sources, columns are testing targets. Scores are the Expected Rank (ER), lower scores are better.

Generalisation of learning models

We next perform re-identification using two learning methods: BR-SVM [6] and attribute learning [101], each of which provides at least near state-of-the-art performance when applied within a single domain. To evaluate cross-domain generalisation, we train the methods on each domain (VIPER, PRID, GRID, CUHK) and apply them to all domains, thus obtaining 16 conditions[3] per method as shown in Table 5.2. The important points to note here are that (i) for both learning methods, the within-domain performance (diagonal of the table) is **significantly** better than the across-domain performance, i.e., *the methods do not directly generalise across-domain*; and (ii) the performance of the learning methods when applied across-domains is actually worse than the low-level feature methods (Table 5.1). This shows that achieving a useful level of performance with learning methods outside of closed-world benchmarks is non-trivial, and hence highlights the value of our contribution in this chapter.

The above results together show that neither low-level features nor learning methods generalise directly and reliably across-domains, therefore the only viable route to good performance *is*

[3]Except for ATTR for CUHK because we had no attribute annotation for this domain.

to learn a new model for each pair of cameras. However, the quadratic number of pairs means that in practice exhaustive annotation is unreasonable beyond benchmark dataset testing exercises. This is turn shows the value of our contribution of transferring re-identification models for reducing training data requirements.

Computational Efficiency

The practically relevant aspect of performance is online matching speed. As a SVM approach, our model is linear in the number of support vectors at test time. In particular it requires S times the computation of [6] for S source domains. In practice this means that our multi-kernel matching took about a millisecond per comparison (79ms including ELF feature extraction) with our unoptimised Matlab implementation. We note that despite making use of a strong model, this is still faster than state of the art LLFs such as SDALF [47], which requires approximately 460ms per comparison.

5.4 Discussion

In this chapter we introduced the problem of domain transfer for re-identification. This is a highly relevant challenge for taking re-identification out of closed-world benchmarks and making it useful for real-world applications. By formulating domain-transfer re-identification as a SVM multi-kernel learning problem, we were able to achieve good performance on a wide variety of public benchmark datasets with a fraction of the training data required by previous methods. Moreover, our approach is able to evaluate available source domains automatically, weighting the relevant sources appropriately and ignoring irrelevant sources, thus avoiding negative transfer. We achieved these results despite the fact that the datasets used were unrelated and independently collected. With a wider selection of source datasets to choose from, the ability to construct a mapping to the target domain of interest (Figure 5.2 on page 127) will be increased [83], and our results are therefore expected to only improve further as additional datasets are released.

There are many remaining opportunities for future work to improve upon the methods explored in this chapter, primarily we wish to further reduce the amount of training required data whilst maintaining good performance in the target domain. Additionally, we have only used the simplest colour feature available in this chapter; absolute performance should improve when using "better" features as input, and multiple different features can readily be incorporated into our MKL framework. With regards to negative instance selection, we thus far randomly selected 10

VIPeR, Rank:	1	10	20	50	**PRID**, Rank:	1	10	20	50
BR-SVM 1/2	12.34	55.22	74.37	91.77	BR-SVM 1/2	3.00	19.00	35.50	76.50
DTR 1/2	13.45	51.58	74.68	92.72	DTR 1/2	2.50	21.50	36.00	74.50
BR-SVM 1/4	6.49	42.72	62.66	86.71	BR-SVM 1/4	2.50	15.00	31.00	69.00
DTR 1/4	9.02	45.09	68.20	88.29	DTR 1/4	2.50	14.50	30.50	68.50
BR-SVM 1/8	5.06	30.38	47.78	72.78	BR-SVM 1/8	2.00	15.00	30.00	58.00
DTR 1/8	5.85	34.49	53.96	78.32	DTR 1/8	2.00	15.00	28.50	58.00
BR-SVM 1/16	3.48	22.15	37.50	58.86	BR-SVM 1/16	1.00	15.50	26.50	56.50
DTR 1/16	5.54	27.06	44.15	67.41	DTR 1/16	1.00	17.00	27.00	58.50
GRID, Rank:	1	10	20	50	**CUHK**, Rank:	1	10	20	50
BR-SVM 1/2	14.40	58.40	76.40	96.40	BR-SVM 1/2	7.84	33.09	45.88	68.25
DTR 1/2	18.00	62.80	79.60	96.00	DTR 1/2	8.04	34.64	47.73	67.01
BR-SVM 1/4	12.00	52.40	66.80	87.60	BR-SVM 1/4	5.46	27.22	40.21	59.90
DTR 1/4	16.40	53.60	72.80	93.20	DTR 1/4	6.80	29.90	42.89	63.61
BR-SVM 1/8	6.80	41.60	64.80	89.20	BR-SVM 1/8	3.30	17.32	26.60	46.49
DTR 1/8	12.40	47.20	66.80	91.60	DTR 1/8	5.46	20.52	33.61	52.78
BR-SVM 1/16	6.00	30.80	49.60	77.60	BR-SVM 1/16	1.96	8.97	16.49	32.99
DTR 1/16	8.00	37.60	55.20	81.60	DTR 1/16	2.89	13.71	23.20	40.41

Table 5.3: We present rank scores for each of the target datasets and annotation volumes for both Binary-Rank SVM (BR-SVM) and our Domain Transfer Re-identification (DTR) approach. Higher scores are more desirable, as are earlier ranks which are more useful to human operators. Our approach shows that even with extremely reduced annotations on the target dataset, re-identification knowledge can be transferred in order to improve performance over low-level features alone.

negative pairs per positive pair for training. Re-identification accuracy can be increased at the cost of additional computation by increasing this ratio [6]. However, more interesting is investigating active learning or instance mining approaches to optimally select the right instances from the quadratic number of pairs is therefore an important open question. Finally, we would also like to transductively exploit the unlabelled data distribution in the target domain, and eventually move towards completely annotation free transfer learning for re-identification. The work from Chapters 3, 4, and 5, all share an underlying and important assumption with all other re-identification work to date; namely that our surveillance cameras are statically fixed in place. In the final technical chapter, we investigate what setting aside this assumption means for next-generation, real-world, re-identification systems.

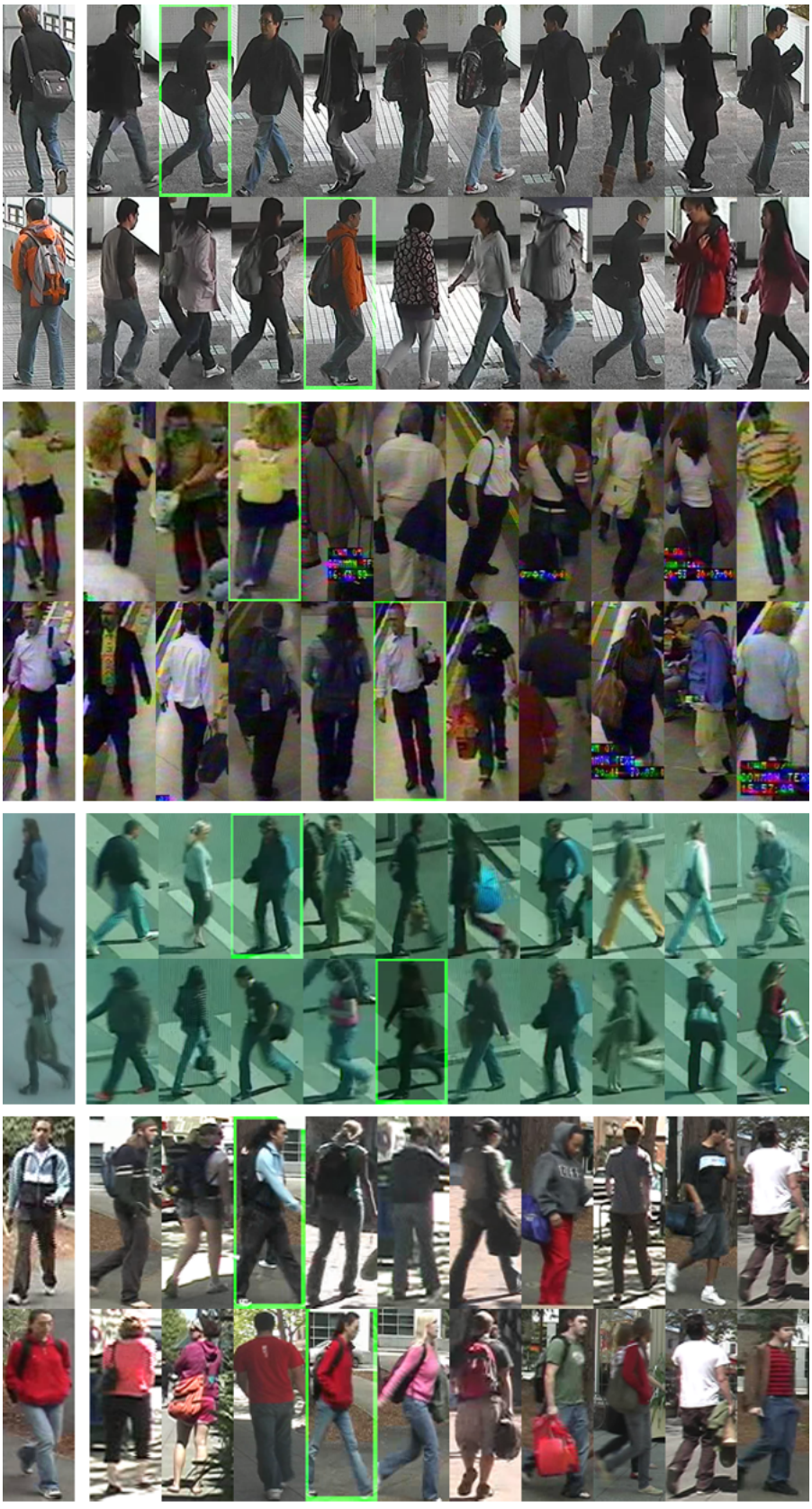


Figure 5.7: Some examples of early-rank matches from our system. The leftmost image is the probe image, with gallery images ranked by similarity to the right. The correct match to the probe is highlighted in green. From top to bottom, we present two examples from VIPeR, PRID, GRID and CUHK.

Chapter 6

Exploring the Open World

A fundamental assumption in almost all existing re-identification research is that cameras are in fixed emplacements, allowing the explicit modelling of camera and inter-camera properties in order to improve re-identification. In this chapter, we present an introductory study pushing re-identification in a different direction: re-identification on a mobile platform, such as a Web. We formalise variants of the standard tasks for re-identification that are more relevant for mobile re-identification. We introduce the first dataset for mobile re-identification, and we use this to elucidate the unique challenges of mobile re-identification. Finally, we re-evaluate some conventional wisdom about re-identification models in the light of these challenges and suggest future avenues for research in this area.

6.1 Problem Definition

Person re-identification has been extensively and aggressively studied in recent years by the computer vision community due to its challenging nature and critical role in underpinning many security and business-intelligence tasks in multi-camera surveillance [65]. This has resulted in continued improvements in performance on increasingly challenging benchmark datasets. In essence, re-identification is about successfully retrieving people by *identity*, enabling security operators or higher-level software components to locate individuals. Nevertheless, it is conventionally formulated as a one-to-one set-matching problem between two fixed cameras, for which an effective model can be learned. In this chapter we present an introductory study that relaxes this core assumption and investigates how re-identification generalises to mobile surveillance

platforms as realised by unmanned aerial vehicles (UAVs) [34].

Despite the successes of static CCTV cameras, we argue that considering alternative surveillance equipment not only opens up exciting new research areas, but also new ways of thinking about re-identification and particularly, how re-identification fits into real-world applications and links with other research fields. New technology such as remotely-operated vehicles and wearable visual sensing equipment is becoming increasingly accessible in terms of cost and availability to the general public. In many cases, quickly deployable mobile visual systems rival currently predominant static CCTV cameras in terms or resolution and frame-rate. More critically, they intrinsically have a qualitative flexibility advantage – in terms of being mobile – and are thus able to dynamically adapt their viewing position and direction without being constrained by the emplaced locations of a CCTV camera. We term any piece of equipment that can be exploited for the acquisition of video data for surveillance – and particularly in a portable sense – a *mobile re-identification platform* or, MRP.

Generalising re-identification to MRPs provides many new capabilities and research avenues, as well as introducing some significant differences and new challenges compared to the standard formulation of the re-identification problem. These broadly relate to the interrelated issues of (1) view ambiguity, (2) view variability and (3) open-world re-id.

6.1.1 Within-view Ambiguity

The first major contrast between MRP and standard fixed camera re-identification relates to the number of views. That is, the standard setting is typically defined across a pair of camera views, and within-camera tracking is typically assumed to fully disambiguate detections within-view. In contrast for MRPs 'within camera' re-identification is itself non-trivial because the camera's positional and orientational mobility means that even stationary people frequently enter and exit the view area due solely to self-motion of the platform. This further generalises the so called 'M vs All' scenario described in [87] to 'All vs All'.

6.1.2 View Variability and Generality

The second major contrast is the continually varying view-stream of a MRP compared to the conventional fixed position CCTV camera. This is significant because most of the recent performance gains in the state of the art re-identification methods have come from supervised learning of *view* or *view-pair specific* models [66]. In the MRP case the *continually* varying view param-

Figure 6.1: Comparison of typical surveillance scenes from (left column) standard surveillance data from a static CCTV camera and (right column) from a mobile re-identification platform (MRP). Examined side-by-side, it is clear that the CCTV camera footage is more suitable for discriminative machine learning as the variability of the human detection appearances are at least somewhat constrained for emplaced cameras; this assumption is dramatically violated in the case of MRPs since MRPs are much less constrained, therefore relative viewpoint variance is more pronounced.

eters – including range, lighting, self induced motion blur and detection alignment – precludes learning such models (see Figure 6.2 on the facing page).

The conventional approach to maximising re-identification performance is learning a discriminative model to maximise re-identification rate for a specific pair of fixed camera views [94, 6]. A few studies have started to consider how re-identification models generalise across views [103] and generally found that achieving good re-identification rate requires *view specific discriminative training*. Specifically, that camera view covariates must be learned individually and that each camera view requires individual training in order to train a good model. This reflects analogous conclusions drawn more broadly in computer vision recognition [170]. As a result, studies have begun to develop transfer strategies that allow models learned from 'source' view pair(s) to be adapted to better apply in a new 'target' view [25, 103, 121] which may have different position, lighting, etc. These studies have generally considered combining [25] or adapting [103, 121] source model(s) to construct the model for a new domain – with the general aim of reducing or eliminating the need for collecting annotated training data for every pair of cameras.

The important contrast with our MRP context is that domains, or camera pairs, as described above are *discrete*. In contrast, the video feed from a MRP is a *continuously varying* domain. This means that for previous approaches to view generalisation it is still assumed that enough data to model a specific view or view pair can be collected and a discriminative model learned. This is no longer feasible for MRP, since the constantly varying view means that collecting (let alone annotating) extensive view-specific data is impossible, and the conventional strategy of learning a discriminative model is called into question.

6.1.3 Open-world

Most existing re-identification studies make the simplifying assumption of closed-world conditions. That is, there is a one-to-one set match, where everyone in the first camera re-appears in the second camera. No one disappears, and no extra people appear. Although convenient for modelling and benchmarking purposes, this is clearly an extremely strong assumption in practice. In the case of MRP with within-camera re-identification ambiguity, and the mobile nature of the platform, closed-world is clearly an inappropriate assumption – meaning that re-identification with MRP is significantly more ambiguous than the conventional setting.

At its most general, open world re-identification [65, 27] addresses relaxing several assumptions: one-to-one set-match (that is, that every person in the probe set appears in the gallery set

Figure 6.2: Illustrating key differences in person detection quality when automatically detected from mobile re-identification platform video (MRP, left), compared to detections in a standard re-identification dataset, VIPeR (right). Notably, the VIPeR images (i) are in perfect register, (ii) feature standard walking poses from a limited number of relative angles. Contrastingly, the MRP images are unregistered, feature more varied pose and also occasionally heavy motion-blur because of the relative motion of the MRP to the target person during transit.

---

and vice-versa) [84]; the assumption of matching between only two cameras [84]; the assumption of a known number of people; or that multi-shot grouping is known a-priori [87]. A few studies have begun to work toward this including [84, 87]. However, these have generally considered only a couple of these relaxations at once. In contrast, the MRP re-identification scenario is intrinsically open-world: self movement in a potentially open-space means one-to-one match situations are unlikely, self-motion means that tracking cannot provide multi-shot grouping, and clearly the person count of an arbitrarily surveilled space is not known in advance.

Despite the challenges identified above, MRPs provide a compelling new ground to break for re-identification science both in terms of broadening the application area as well as providing the opportunity to reconsider several implicit but strong assumptions made in most existing re-identification research. In this work, we make four main contributions: (i) We present a case for the pursuit and development of a new research area using mobile re-identification platforms (MRPs); (ii) We formalise three novel MRP-related variants on the classic re-identification scenario; as well as associated evaluation metrics for each; (iii) We collect the first public dataset for MRP re-identification and establish benchmarks for each of the identified tasks; (iv) We elucidate the unique challenges posed by MRP re-identification and discuss their implications for general re-identification research going forward.

Going beyond conventional re-identification, we next discuss a few recently identified research areas that are relevant to our MRP context.

6.1.4 UAVs

A full discussion of background research in UAV technology is out of the scope of this chapter, but see [34] for an introduction and background to UAVs and their capabilities. The central

issue for UAVs to become more useful for surveillance tasks is for them to become increasingly autonomous, and a significant component of this is learning to maintain consistent person identity estimates over time, which we address here.

6.2 Re-identification Problem Variants and Metrics

Conventional re-identification is used as a forensic search tool, or as a module by higher-level software – such as inter-camera tracking [149]. For ease of model formulation (e.g., metric learning, SVM ranking), evaluation and establishing benchmarks, most studies formalise re-identification as a closed-world set match between two specific cameras. As a result the typical evaluation metric is Rank 1 accuracy (the % of perfect gallery matches for each probe image), or the CMC curve (the % of correct matches within the top N ranked matches, for varying N) [178]. In this section we describe three distinct variants of the re-identification problem that naturally arise with MRPs – each based on intuitive application scenarios for a MRP. Table 6.1 on page 151 summarises the problem variants proposed and compares them with classical approaches to re-id.

6.2.1 Watchlist Verification

In the *watchlist* task, the MRP is patrolling an area and the goal is to detect if any person encountered is somebody on a pre-defined watch-list. For the moment we make no assumption on whether the MRP is manually controlled, has a pre-programmed travel path or autonomously wanders. However, we assume that the scenario is *passive sensing* – the MRP is not going to to take action based on any detected matches. The watchlist itself could come from a variety of sources: a pre-defined mug-shot gallery; a transmitted detection from another MRP or CCTV camera; or a previous detection saved by the current MRP on a previous flight or earlier in this patrol. For example the MRP may be trying to track down a specific person previously identified performing a suspicious action of interest.

In this case, the 'probe' is a single person from the watch list, and the 'gallery' is all people observed in a patrol (see Figure 6.4 on page 150). In contrast to conventional re-identification (see Figure 6.3 on the next page), this is a more open world problem in that: (i) the probe person may not appear anywhere in the patrol video (no match is an option), (ii) (most) people in the patrol video are not on the watchlist (many background distractors), and (iii) the total number of detected instances of the true match if present in the gallery/patrol video is unknown (not one-to-

one). In Table 6.1 on page 151 this is illustrated under match by [N] and [M] reflecting multiple potential *ungrouped* matches and distractors respectively.

Given these considerations, the right evaluation metrics for this problem are information-retrieval style metrics, thus we use a suite of them: (i) the rank of the true matches, and (ii) precision-recall curves and associated summary – average-precision.

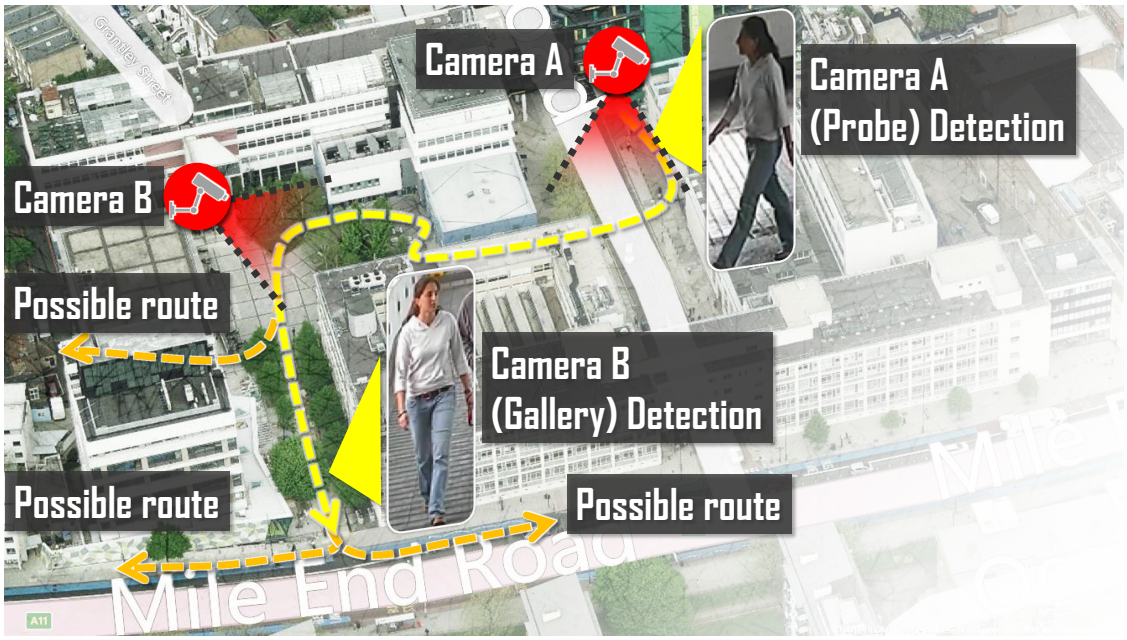


Figure 6.3: Illustrative example of a real-world re-identification set-up using static cameras, the type of scenario which most re-identification work assumes. A person travels across an urban public space (yellow path), and may take multiple potential routes whilst passing between CCTV blind spots (orange paths) where they cannot be detected and their location is therefore unknown. Contrast this scenario with that in Figure 6.4 on the following page where a mobile re-identification platform may maintain surveillance on the target throughout the entire path, as well as potentially follow the target if they deviate from an expected route.

6.2.2 Within-Flight Re-identification

In the *within-flight* re-identification task, the MRP's goal is to maintain consistent identity of person detections recorded throughout the flight. Due to both platform and target motion, a particular target may enter the view once, or enter and exit the view multiple times throughout the flight. In this case there is only one "camera view" as compared to conventional re-identification setting of two fixed cameras. However, it means that: (i) the platform motion can create potentially more view-variation over time than occurs between two fixed CCTV cameras, so "within-view" re-identification can become even harder than conventional re-id; (ii) as before, there is a general

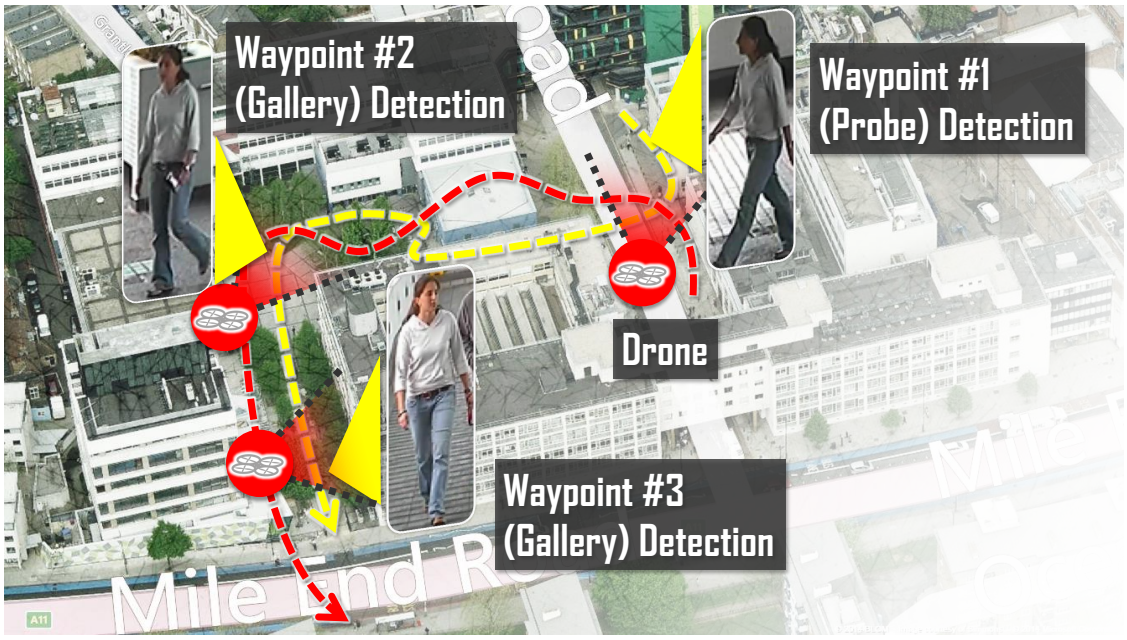


Figure 6.4: Illustrative example of a real-world re-identification set-up using a mobile re-identification platform (MRP), or UAV. Contrasting with the same scenario using static CCTV (Figure 6.3 on the preceding page), the UAV is able to follow (red path) the path of the person (yellow path). This enables surveillance in areas with no existing CCTV coverage, but at the cost of person detections that are more diversified in terms of their appearance and their appearance under the effect of practical covariates such as viewing angle and lighting conditions.

---

open-world identity inference problem.

The general identity inference problem here means that there is no-longer a notion of probe and gallery. Instead there is a list of N detections, to which we wish to assign one of $K \leq N$ unique identities for later tasks to use. However K (the number of unique people in the scene) is itself unknown. In Table 6.1 on the next page this is illustrated under match by [N] – the single set of detections with unknown grouping – and an unknown person count.

Evaluating this open world identity assignment is non-trivial compared to closed world. To fully evaluate the performance, we use statistical analysis on all pairs of detections to measure pairwise Precision and Recall. Specifically given all true $\mathcal{L}_{gt}$ and estimated $\mathcal{L}_{est}$ labels of the N detections. A 'true' pair i, j has the same label, and a 'false' pair have different labels. Thus true-positive, true-negative, false positive and false-negative rates can be computed as in Equation (6.2 on the facing page); which can in turn be summarised in terms of Precision, Recall, Specificity, and Accuracy as in Equation (6.1 on the next page).

Setting	Cameras	Match	Person Count	View-specific	Multi-shot	Evaluation
Singleshot [47, 186, 94, 6]	2	$N:N$	Known	Yes	No	Rank 1, CMC
Multishot [47, 91]	2	$N:N$	Known	Yes	Grouped	Rank 1, CMC
Karaman [87]	2	$N:[N]$	Known	Yes	Group : No group	Accuracy
John [84]	2	$N+M_1:N+M_2$	Known	Yes	No	Rank 1
Watchlist	1	$1:[N]+[M]$	N/A	No	No group	Rank, Prec+Recall
Within	1	$[N]$	Unknown	No	No group	F-measure
Across	2	$[N]+[M_1]:[N]+[M_2]$	Unknown	No	No group	F-measure

Table 6.1: Contrasting re-identification problem variants. Match: $N:N$ reflects closed world one-to-one mapping among N people in view 1 : view 2. $[N]$ indicates unknown within-camera grouping. M represents the unknown fraction of the people to be matched who are distractors in that they do not occur in the other view or the watchlist.

$$\begin{aligned} Prec &= TP/(TP+FP) \\ Rec &= TP/(TP+FN) \\ Spec &= TN/(FP+TN) \\ Acc &= (TP+TN)/N \end{aligned} \tag{6.1}$$

$$\begin{aligned} TP &= \sum_{ij} (\mathcal{L}_{gt}(i) = L_{gt}(j)) \wedge (\mathcal{L}_{est}(i) = \mathcal{L}_{est}(j)) \\ TN &= \sum_{ij} (\mathcal{L}_{gt}(i) \neq L_{gt}(j)) \wedge (\mathcal{L}_{est}(i) \neq \mathcal{L}_{est}(j)) \\ FP &= \sum_{ij} (\mathcal{L}_{gt}(i) \neq L_{gt}(j)) \wedge (\mathcal{L}_{est}(i) = \mathcal{L}_{est}(j)) \\ FN &= \sum_{ij} (\mathcal{L}_{gt}(i) = L_{gt}(j)) \wedge (\mathcal{L}_{est}(i) \neq \mathcal{L}_{est}(j)) \end{aligned} \tag{6.2}$$

6.2.3 Across-flight Re-identification

The *across-flight* problem is somewhat more related to the classic problem of between-camera re-id. In this case identities should be matched across two separate MRP flights. This may be from either the same platform making two patrols, or two distinct and communicating platforms trying to coordinate identities. It is a fully open-world problem, given that within-flight/view tracking cannot be assumed for MRPs (ungrouped detections in Table 6.1), and that only an unknown subset of the total people in each view may be shared (in Table 6.1, N shared + M distractor people in each view). However, compared to within-flight re-identification, it may be somewhat harder because the environments across space and/or time may be even more different than the view change caused by platform motion in the previous case. Again, statistical analysis is the appropriate evaluation technique.

6.3 Methodology

6.3.1 UAV Setup

We use a retail remote-operated quadrocopter to realise our MRP for the purposes of data acquisition (see Figure 6.5 on page 153). During data collection, a human operator pilots the UAV via

laptop using the Robot Operating System (ROS[1]) to ensure responsive handling with the control loop; sensor data capture was performed in parallel and at ≈ 200Hz whilst video from the quadrocopter was sampled at $\approx 1-5$Hz. For this particular commodity platform, flight time was limited by UAV platform weight (436g) and battery capacity to $\approx$ 10 minutes per flight. The UAV possesses two cameras, of which only the main camera is used. This camera is a diagonal lens, CMOS camera providing a 90°field of view at a theoretical maximum quality rating of 1280x720 (720p) and 30 frames per second (fps). Because of experimental considerations, such as the computational processing required to generate the real-time person detections, the video was recorded at 640x360 pixels (*i.e.* subsampled 50%). Although this particular retail UAV has a top speed limit of 18km/h, due to safety considerations and the goal of acquiring optimal video data for person detection and re-identification, the UAV's maximum lateral velocity was constrained to little more than normal human walking speed and the maximum flight "ceiling" set to an altitude of 15 meters. Finally, in order to compensate for environmental factors affecting human control, the UAV employs an ARM cortex A8 CPU operating at $\approx$ 1Ghz to provide stabilisation assistance for the pilot.

During flight, a heads-up-display (HUD) is overlaid on top of the video feed displaying standard sensor information (such as yaw, pitch, acceleration, battery and altitude), as well as real-time person detections and person detection confidence scores. This in some sense serves to provide the operator with the visual cues necessary to weakly simulate an active-sensing, fully autonomous (i.e. closed-loop) UAV. If the UAV is orientated poorly towards a person or the person is partially occluded then a poor detection will result and the operator can adjust the relative orientation and position of the UAV based on this visualisation until a strong detection can be obtained. Some examples of the HUD can be seen in Figure 6.6 on page 154.

6.3.2 Person detection

Given the $1-5$Hz video feed, the next task is to obtain person detections. To maximise the reliability of this step, we first apply a corrective transform on each frame to correct for the 'roll' of the UAV (using data recorded from the MRP's onboard accelerometer sensor), since the detection models assume people to be upright. In order to detect people fast enough for real-time visualisation so as to assist the MRP's operator, we employ [43]'s toolkit which provides excellent computational efficiency and detection quality. At extraction time, we resample detections to

[1] http://www.ros.org/

Figure 6.5: Photographs showing in-flight detail of the retail UAV used during our data capture sessions. Since the main camera is mounted on the front of the UAV (see bottom-right), lateral motion ("strafing") or rotation ("panning") wass employed to direct the camera at new targets; however, this can result in motion blur and introduces a new challenge unique to this surveillance modality.

Figure 6.6: Illustrative examples of our mobile re-identification platform's human interface as used in the data capture sessions; illustrating real-time person detections colour-coded by detection confidence. The top-left and top-right images illustrate typical operator views from the outdoor and indoor flights from Dataset 1; The bottom row illustrates Dataset 2. See Figure 6.7 on the facing page for a description of graphical components.

[128x48] pixels[2]. We threshold detections and discard any with a confidence of below 20% since the environments from which we will be detecting are extremely varied with respect to lighting and pose and we wish to limit the number of potential false-positive detections whilst retaining most true detections. For our visual features we employ the commonly used ensemble of local features (ELF) [68], which encodes both colour and texture in 6 horizontal strips [147] for final features of 2784 dimensions.

6.3.3 Datasets

Using the procedure described above, we collected two multi-flight datasets. The first dataset contains three flights worth of data, across an outdoor and indoor environment. These consisted of 436, 652, and 848 video frames, from which we obtained 233, 471, and 797 person detections from 6, 7, and 10 distinct people (after thresholding). All person detections in this dataset are exhaustively annotated.

The second, significantly larger, dataset contains six flights of data in three different unconstrained and heavily crowded outdoor environments. Across each flight there are between 10,000

[2]However, note that the original resolution and therefore resample quality will vary dramatically over time within a flight, see Figure 6.1 on page 145

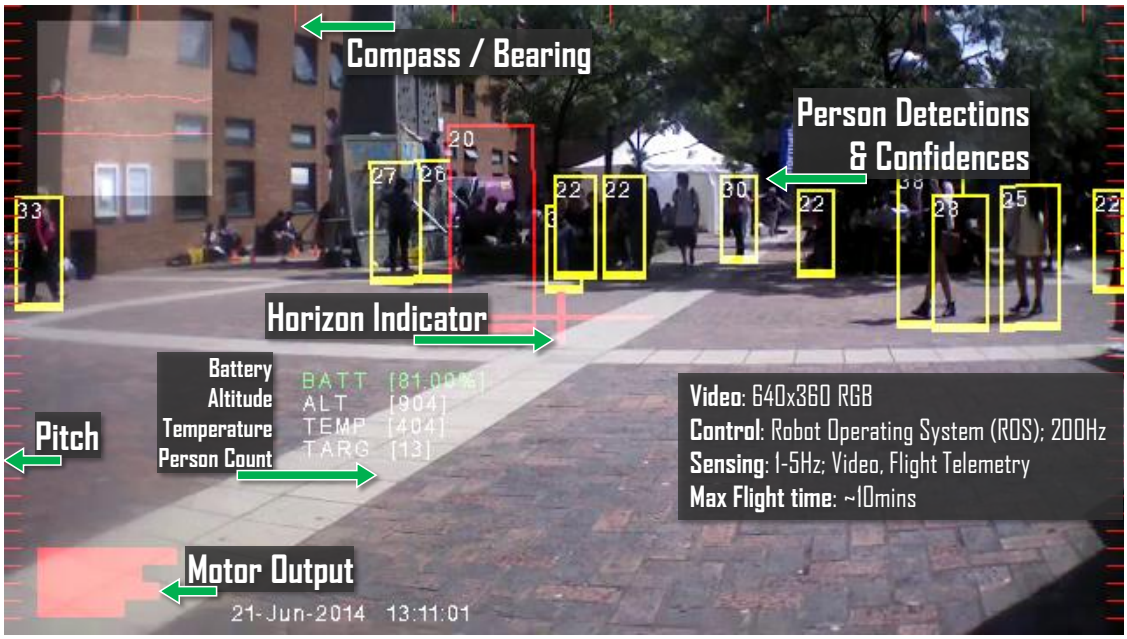


Figure 6.7: Anatomy of the heads-up-display (HUD) used by the UAV operator to simulate the visual cues used in a closed-control-loop mobile re-identification platform (MRP). The HUD overlays real-time person detections over humans and uses colour coding to indicate detection confidence as well as reporting qualitative confidence scores and other standard telemetric readings to assist with piloting. The human-detection bounding boxes assist the pilot in manoeuvring the UAV for optimal human detection (*i.e.* minimal misalignment of detection bounding boxes and therefore cropped person images at the downstream re-identification step.).

---

and 30,000 frames of video data and an average of 8,654 person detections from an unknown number of distinct people. Of this data, we selected a single flight and exhaustively annotated 28 unique identities within the 4096 detections available within a 2:06 window.

6.3.4 Classifier training, Representation and Datasets

One of the central questions we wanted to answer is to what extent the state of the art discriminative models for standard benchmark datasets are effective for MRP based re-identification. This question is crucial because conditions in MRP-sourced video data continuously change during a flight thus there are many more combinations of pose and viewing angle than in the fixed view case assumed by most state of the art models – i.e. a fixed view with enough (annotated) data is sufficient to learn a model. It is therefore critical to discover if and how much performance discriminative models lose on dynamically changing data.

We investigate this by training a selection of strong discriminative models including one of the most popular: RankSVM [147]; and two recent state of the art approaches BR-SVM [6] and

KISS [94]. We train these models on a variety of large benchmark datasets including VIPER [68] (632 distinct persons in [128x48] crops), PRID [75] (200 distinct persons), GRID [120] (250 persons) and CUHK [109] (971 persons). We resample all detections to match VIPeR's dimensions. For the computationally intensive discriminative methods, we reduce the dimension with PCA to $d = 200$ for BR-SVM and $d = 34$ for KISS as specified in [94].

6.3.5 Domain Shift

Since we assume a stationary view and the absence of live-annotation of video-feed data (as proxies for normal discriminative training on a single-view), the only way to apply trained matching models for MRPs is to train them on benchmark datasets before testing them on the MRP video feed. This potentially opens up the issue of *domain shift* [64, 140, 103] between the training and testing data. For example, due to additional chance of motion blur, mis-registered images and more variance in pose from the MRP detections (Figure 1), which are absent in VIPER.

As a preliminarily investigation into how to overcome this issue, we consider unsupervised domain-adaptation in order to better align the target MRP data X_t and source VIPeR training data X_s. That is, warp $p(X_t)$ so that it is more aligned with the source training data $p_{adapt}(X_t) \approx p(X_s)$, with the intuition that this should allow classifiers trained on X_s to generalise better to X_t [140]. In particular, we align the projected subspaces of the two datasets, using the geodesic flow kernel domain adaptation (DA) method [64] using $d_{DA} = 13$ dimensions.

Intuitively explained, the process of alignment involves treating the subspaces of domains X_t and X_s as points on a Grassmannian manifold $\mathcal{G}(d,D)$. The manifold itself can be interpreted as a geometric representation of how imbricated the underlying distributions of features are, and thus the distance between X_t and X_s on this manifold can be viewed as a measure of similarity between the covariate properties inherent to X_t and X_s. Calculating the geodesic flow permits the parametrisation of how the source model transitions (t) to the target; $t = 0$ indicating that a particular projection ϕ is unlikely to be near the target domain, and $t = 1$ indicating high likelihood of being close to the target domain. With the full set of $\mathcal{T}$ subspaces, a kernel may be computed that describes this transition and through which the optimal alignment projection may be found by greedily searching the best number of feature dimensions that result in X_t and X_s being proximal and thus promoting more uniform downstream classification performance.

6.3.6 Re-identification Baselines

For **Task 1: Watchlist**, we simulate this experiment by taking each person detection in turn as the watch-list, and matching it against every other detection from the flight to produce a ranked list. The ranked list of results is then evaluated for relevance with information retrieval metrics (Sec 6.2.1). Whether first, average or last rank; or average precision is the most relevant metric will depend on the end-user application and cost function. We evaluate this task with both Datasets 1 and 2.

For **Task 2: Intra-flight re-identification** and **Task 3: Inter-flight re-identification** (see Sec 6.2.2-6.2.3), the experiment is performed by matching every detection against every other detection. The resulting detection-affinity matrix is thresholded[3] and analysed for connected components [167]. Each connected component defines an estimated person. The estimated $\mathcal{L}_{set}$ and true $\mathcal{L}_{gt}$ identities are compared using statistical analysis as explained in Section 6.2. We evaluate these tasks with Dataset 1. As algorithms to produce the matching scores for each experiment, we compare the following models:

NN-[DA] Nearest-neighbor (NN) matching based on the detection descriptor.

BR-SVM-[DA] Binary-relation SVM with RBF concatenation kernel [6].

RankSVM-[DA] SVM with difference feature and linear kernel [147].

KISS-[DA] State of the art discriminative Mahalanobis metric learning [94].

In each case we compare the model with and without domain adaptation (-DA suffix). As explained earlier, we do not have annotated view-specific training data. Thus, we train the latter three discriminative models on the full VIPER dataset of 632 pairs and test them on the MRP video detections. These models obtain good results when applied within-domain on VIPER [6, 147, 94], however our experiment will test their ability to generalise this knowledge to a continuously varying view.

6.4 Experiments

6.4.1 Watchlist and Re-identification Evaluations

We first present the results for the three main tasks before drawing conclusions from them.

[3]The threshold is chosen to optimise F-measure for each model.

The results of watchlist verification are presented in Table 6.2 on the next page (top) for Dataset 1, and Table 6.2 on the facing page (bottom) for Dataset 2. This task reflects how highly true matches to each particular watchlist person are ranked relative to all the other person detections in the dataset, on average. Clearly all methods perform better than random: average rank, for example, has a chance level of half the number of detections across all flights which is $500/2 = 250$ for Dataset 1 and $4046/2 = 2023$ for Dataset 2. The best methods obtain a first rank result of around 2. Surprisingly, this is the case both in the smaller Dataset 1 and the larger Dataset 2.

Intra-flight re-identification results for Dataset 1 are presented in Table 6.3 on page 160 (top). This task attempts un-constrained detection association across all detections within a flight.

Intra-flight re-identification results for Dataset 1 are presented in Table 6.3 on page 160 (bottom). This task attempts un-constrained detection association across all detections from a pair of flights.

6.4.2 Observations and Analysis

Based on the results described in the previous section and Tables 6.2 on the next page- 6.3 on page 160, we make the following observations and conclusions.

(1) NN is best overall – Surprisingly, outperforming all discriminative methods including KISS, BRSVM and RankSVM. In dramatic contrast to the standard ordering of results obtained in the literature [94, 6, 147], where discriminatively trained models significantly outperform simple nearest-neighbour; our results show that in the MRP context, the simplest NN method is generally best. This is true overall for Dataset 1 with all three tasks, as well as the significantly larger Dataset 2 for the watchlist task. This is due to the intrinsic challenge of MRP re-identification that there is no possibility to learn view-specific models.

In order to apply discriminative models to our MRP data, we transferred models trained on VIPER. However, this may not be effective because the MRP video is more variable and unconstrained. Meanwhile, the strong discriminative models have evidently over fitted to the more constrained viewing conditions in VIPER. NN, in contrast, is more reliable because it doesn't

Dataset 1	NN	NN-DA	KISS	KISS-DA	BRSVM [6]	BRSVM [6] DA	RankSVM [147]	RankSVM [147] DA
First rank ↓	2.08	4.69	4.15	5.37	12.32	15.87	9.76	17.53
Last rank ↓	167.93	162.89	156.70	150.82	166.35	160.78	177.32	170.37
Average rank ↓	56.30	56.47	54.45	57.39	65.65	68.77	76.81	81.51
Average Prec ↑	0.46	0.46	0.43	0.41	0.34	0.35	0.24	0.24

Dataset 2	NN	NN-DA	KISS	KISS-DA	BRSVM [6]	BRSVM [6] DA	RankSVM [147]	RankSVM [147] DA
First rank ↓	1.91	2.47	18.02	9.89	265.59	18.87	280.57	424.64
Last rank ↓	1864.34	2001.95	2152.18	2032.83	2841.16	2238.77	2673.06	3357.40
Average rank ↓	507.30	528.85	619.78	635.10	1256.77	753.53	1213.27	1848.23
Average Prec ↑	0.36	0.34	0.19	0.25	0.04	0.14	0.04	0.02

Table 6.2: Watchlist verification results for each model. Top: Dataset 1, results are averages over all persons and all flights, average 500.3 total detections. Bottom: Dataset 2, results are for single annotated flight, 4046 total detections. For the rank metrics lower is better (↓) and for the average precision metric higher is better (↑).

	Precision ↑	Recall ↑	F-Measure ↑	Specificity ↑	Accuracy ↑	Precision ↑	Recall ↑	F-Measure ↑	Specificity ↑	Accuracy ↑
NN	**0.83**	0.29	**0.39**	**0.99**	**0.88**	**0.34**	0.49	**0.29**	**0.63**	**0.60**
NN DA	**0.47**	**0.59**	**0.47**	0.76	0.73	**0.38**	0.39	**0.32**	**0.80**	**0.74**
KISS [94]	0.32	0.30	0.28	**0.82**	**0.74**	0.15	0.93	0.26	0.09	0.21
KISS [94] DA	0.23	**0.59**	**0.31**	0.56	0.56	0.15	0.97	0.26	0.04	0.18
BRSVM [6]	**0.37**	0.27	0.18	0.79	0.70	0.15	**1.00**	0.26	0.00	0.15
BRSVM [6] DA	0.32	0.23	0.17	**0.85**	**0.74**	0.15	**1.00**	0.26	0.00	0.15
RANKSVM [147]	0.00	**0.65**	0.17	0.35	0.38	0.15	**0.98**	0.26	0.03	0.17
RANKSVM [147] DA	0.00	0.36	0.12	0.64	0.58	0.15	**0.98**	0.26	0.03	0.17

Table 6.3: Re-identification results for Dataset 1: (left) Intra flight, and (right) Inter flight. In each case Precision, Recall and F-measure are averaged across all three flights. Higher is better for all metrics, key results are emphasised with **bold-face** font.

train a strong discriminative model and thus cannot over fit in this sense.

(2) Simpler models are better overall The overall ordering of the results is $NN > KISS > BRSVM$. This generally reflects the model complexity, with NN being the simplest. BRSVM being the most complex (due to RBF kernels on concatenated data), and KISS being in between. This ordering also reflects the importance of pairwise training data volume to the model, with KISS and BRSVM both requiring fairly large volumes of training data from the same view in order to perform well.

(3) Domain adaptation can help – but it helps NN significantly more than discriminative models. Comparing the un-augmented condition of each model with the domain adaptation condition (-DA suffix), we see that domain adaptation doesn't make much consistent difference for the watchlist experiment (Table 6.2 on page 159), but it sometimes makes a significant difference in the re-identification experiment (Table 6.3 on the facing page). However, KISS for example is improved from mAP of 0.28 to 0.31 with domain adaptation; while NN is improved much more significantly from mAP of 0.39 to 0.47. That domain-adaptation can help is in one sense not surprising (the MRP video has different statistics to VIPER and aligning the distributions should help), but in another sense surprising (the MRP video is only a *domain* in a very limited sense – because the view varies so much there is hardly a consistent set of statistics $p(X_t)$ to adapt toward). Meanwhile, the fact that it helps NN more than KISS is understandable because KISS still suffers from over fitting to the particular source data (VIPER).

(4) Discriminative models cannot be "fixed" for MRP by adding more conventional training data. The significance of the previous results – with respect to limitations of the discriminative models – could be questioned on the grounds of whether VIPER data is *representative* enough for the variety of views obtained by the MRP. To test this, we re-trained the KISS model using the union of the four largest benchmark re-identification datasets to date, including VIPER, CUHK, GRID and PRID, thus greatly increasing the volume and variety of data used. Table 6.4 on the next page compares the watchlist verification results when training KISS only on VIPER versus training on all existing datasets (ED suffix). Clearly using all the extra data makes only a minor difference to the performance.

	First rank ↓	Last rank ↓	Mean rank ↓	Av Prec ↑
KISS (ED)	1.66	64.44	20.79	0.57
KISS-DA (ED)	3.29	60.68	21.40	0.56
KISS	1.25	81.31	25.90	0.53
KISS-DA	3.50	81.65	30.08	0.35

Table 6.4: Attempting to improve the performance of KISS [94] on the watchlist task by training on all available data (ED). Results are from a single flight in Dataset 1.

6.4.3 Person Count Evaluation

As a final example application, we perform person counting on the flight videos. This is computed as a by-product of open-world re-identification: each identified connected component of the detections defines a distinct person. In general NN and NN-DA provide a near best or best estimate in each case, as seen in Table 6.5.

	Actual	NN	KISS	BRSVM	NN-DA	KISS-DA	BRSVM-DA	RankSVM
Flight1	6.0	±16.0	±23.0	±79.0	±7.0	±20.0	±37.0	±102.0
Flight2	7.0	±0.0	±0.0	±5.0	±1.0	±3.0	±2.0	±2.0
Flight3	10.0	±40.0	±13.0	±1.0	±6.0	±92.0	±3.0	±27.0
Average	7.7	±18.7	±12.0	±28.3	±4.0	±38.3	±14.0	±42.3

	Actual	NN	KISS	BRSVM	NN-DA	KISS-DA	BRSVM-DA	RankSVM
Flight1$\lesseqgtr$2	7.0	±5.0	±0.0	±38.0	±0.0	±0.0	±74.0	±48.0
Flight2$\lesseqgtr$3	10.0	±0.0	±13.0	±21.0	±6.0	±5.0	±0.0	±1.0
Flight1$\lesseqgtr$3	10.0	±0.0	±6.0	±0.0	±3.0	±7.0	±84.0	±226.0
Average	9.0	±1.7	±6.3	±19.7	±3.0	±4.0	±52.7	±91.7

Table 6.5: Person counts in Dataset 1. For each method we report the result as the average error between the estimated and true count. (Lower is better) (upper) **Intra**-flight condition, (lower) **Inter**-flight condition We denote comparisons made *inter-flight* as commutative, with $\lesseqgtr$.

6.5 Discussion

Based on the experiments and analysis in the previous section, we drew the following conclusions: 1. NN is the best method for MRP re-identification, 2. In general simpler methods out-

perform more complex methods, 3. Unsupervised domain adaptation can improve MRP re-id, 4. The challenge is intrinsic to the nature of benchmark datasets being captured by static cameras, and the MRP dataset being captured by a dynamic camera.

Given these observations, we highlight the following considerations for future work:

1. Current re-identification research has been too focused on learning dataset specific models, leading to dataset bias [170]. Analogous to research trends in more general computer vision [92], developing methods that avoid bias and generalise across datasets is necessary to fully exploit the potential of re-identification to MRPs.

2. Domain adaptation methods can potentially help adapt re-identification methods across scenarios with different data statistics. However while most domain adaptation methods require some supervision in the target domain, it is important that DA methods used in this context are unsupervised, since live annotation of MRP detections is implausible. In the current results, a completely disjoint unsupervised DA module [64] is able to make an impact. Investigating tighter integration of the DA and re-identification mechanism is likely to be fruitful.

3. Conventional re-identification and DA [64] methods assume the target task is a distinct and discrete context. The continually varying nature of MRP view, and hence data statistics, means that it may be important to treat MRP as an online rather than a discrete adaptation process. This is a somewhat unique aspect of DA for re-identification in contrast to more general vision problems [170, 92].

4. Consideration of the MRP task highlights the intrinsically open-world nature of re-identification which has largely been ignored for convenience by prior research. In this study we addressed this by a very simple strategy of threshold learning. However, more effort should be put toward developing more systematic and optimal methods to resolve open-world ambiguity.

5. Our new continuously-varying view dataset has a total of 51,922 unconstrained person detections across six flights resulting in hundreds of identities that partially overlap across three outdoor zones. This challenging MRP dataset is qualitatively different to existing re-identification datasets, and will help drive the research challenges identified above.

Finally, given the partial success obtained so far, we discuss some speculative applications for MRP technology.

Our first re-identification case for MRP is an open-loop scenario where the re-identification task does not directly have any impact on the travel path of the vehicle; but data from the vehicle still enables analysis and detection albeit in a passive sense. In this mode of operation, the MRP will likely either be under control of a human operator, or will follow a set of preconfigured waypoints along a patrol-route, with the video sensor data available for analysis either in near real-time, or after the MRP has returned home. This is conceptually closest to the standard re-identification problem.

In contrast, closed-loop MRP control may be fully or semi-automated and critically, may permit the MRP to automatically adapt a regular patrol-route or journey for optimal performance on specific re-identification tasks. For example, re-identification quality-control to move the MRP to get a better view when current re-identification is too ambiguous [152]. For a given flight time or length, this then leads into an interesting trade-off between re-identification accuracy of each individual versus coverage: the fraction of total people captured in a zone in total [162].

Chapter 7

Conclusions

7.1 Goals and Contributions

The primary aims of this thesis has been to explore (i) alternative representations capable of effectively reducing the effect of variations in human appearance after transition to a disjoint camera view in order to facilitate inter-camera entity association, re-identification, (ii) present and explore techniques capable of scaling to real-world use in modern surveillance environments, (iii) review the underlying assumptions that have driven re-identification work in early years, against the recent changes in retail surveillance technology available today.

We adopted an attribute-centric approach to (i) in Chapter 3, developing a mid-level, human-semantic representation that improved re-identification performance, was synergistic with existing features and showed how it can be fused and a mapping function learnt to account for inter-attribute variances in utility and error. For (ii), two methods were investigated, (a) In Chapter 4 a data-driven approach exploiting the copious information available online to discover latent quasi-semantic attributes from meta-text without the need for manual annotation and (b), in Chapter 5 a transfer-learning framework capable of learning inter-camera appearance mappings from multiple camera pairs for transfer to a target domain where less annotation were available. Finally, in (iii) we experiment on a new video surveillance dataset obtained from a retail aerial UAV which violates the traditional assumption that surveillance cameras are statically emplaced in Chapter 6.

In Chapter 3, the attribute representation provides a separate *modality* of feature and therefore

is exploitable for fusion with features from other work such as [47].

The approach of using attributes is advantageous since it produces a lower-dimension feature that facilitates the possible downstream application of normally computationally intensive procedures. Attributes can also be considered a kind of transferable context [190], providing auxiliary information about an instance even when the attribute classifier is not trained on the target dataset. Furthermore, another advantage is the possibility for re-identification by description, such as in the case where one may wish to search for all people wearing "red-shirts" and "dark-pants", or where a visual probe image is unavailable. The attributes we train in Chapter 3 are however, more discriminative when trained on data with proximal practical covariates to the target data and require extensive annotation and the availability of sufficient instances for training on each new camera. The core of this issue is the source of the ontology of attributes – human expert knowledge. Humans can rely on a wide variety of inherently attributes and "soft-biometrics" for re-identification tasks, whereas training modern machine learning discriminative methods to recognise these attributes requires extensive labelling as well as an initial definition of which attributes to annotate. This ontology selection strategy is inherently "top-down" since the human expert defines it according to human intuition without regard to the specifics of the machine learning methodology. This makes it difficult to tell *a priori* whether a given visual attribute (i) can be recognised by a classifier, given (ii) the data available, and (iii) whether it will be informative and useful in discriminating against other people.

In order to alleviate these weaknesses, we investigate a data-driven attribute representation learning framework in Chapter 3. Taking inspiration from Chen *et al.*'s NEIL [30] as well as inspiration from recent data-mining works using the Internet as a source such as Berg *et al.* [18] and Li *et al.* [106] we obtain noisily labelled Internet photographs and their associated meta-text from the Internet using a very broad range of search terms synonymous for "human". Our data are processed in order to build an unsupervised collection of 200 "quasi-attribute" datasets by clustering the information present in the user-defined descriptions for each image's person detections. These form the basis for our unsupervised, Internet-mined mid-level attribute representation, which are composed of latent semantic topics present in the underlying data, such as "paris people" or "camoflage shorts". To verify these clusters successfully encode information as manually defined in the previous chapter, we demonstrate the top retrievals of a regression mapping between the Internet attribute representation from Chapter 4 to the expert attributes as defined

in Chapter 3. This Internet-attribute representation is not as immediately intuitive to humans as the expert-defined attributes however, and requires labelling to learn the mapping function to enable interaction between the two attribute modalities to enable zero-shot re-identification queries. Lastly, the use of LDA classifiers ensures the system scales linearly on the number of attributes required, a necessary requirement to avoid using specialised computation equipment.

Whilst Chapter 4 assumes no labels are available, Chapter 5 considers the scenario where there are some quantity of labelled instances available on the target camera-pair domain, and that other camera-pair domains have been previously trained. In this chapter, transfer-learning is used to learn the nonlinear combination of auxiliary domains that best describe the target domain, given the available labels. The problem is formulated using a multi-kernel SVM model, providing an efficient solver for the complex optimisation task involved and evaluates the source domains automatically whilst learning an appropriate weighting of relevant source domains and simultaneously ignoring unhelpful domains to prevent *negative transfer*. We evaluated this model on public benchmark datasets that were unrelated and disjointly acquired from different locations and times. Despite these differences, our method was successful at discovering only one of the datasets was generally unhelpful, however the others could be assigned positive weightings and contributed to the construction of a target classifier trained on a fraction of the usual label volume required by other methods.

Re-identification research is usually undertaken with a set number of cameras, closed set of probe and gallery images and video from static, immobile camera equipment. Whilst these are reasonable assumptions for many scenarios, recent technological advances have introduced a series of potentially valuable surveillance-capable devices. We term these "MSPs", and in Chapter 6 we formalise some variants of the standard definitions for re-identification that are more relevant for mobile re-identification. These variants are designed to permit investigation into the re-identification paradigm from a different perspective - what happens when the cameras are not statically emplaced? When we don't have entity labels to work with or match together? We conclude that the aggressive pursuit of re-identification research on the limited publicly available benchmark datasets currently available has lead to dataset bias [170], similar to trends in general computer vision research [92] and that relative-pose is an important visual challenge to overcome in future representation research. In keeping with the desire for scalable solutions, a completely disjoint unsupervised DA module [64] is able to make an impact. One particularly critical con-

clusion we draw is that the potential for continuously varying viewpoint change which is inherent to MRP-sourced surveillance data detracts substantially from the performance of previously successful supervised re-identification methods; leaving more basic methods as the best recourse. Lastly, the problem of open-world ambiguity which has until recently been ignored [88, 27], is explored in this context.

Although Chapters 3, 4, 5 and 6 are standalone in that in this thesis we treat them separately, they are also synergistic. They cover many aspects that an all-aspect re-identification pipeline would require for use in real-world, real-scale surveillance applications.

7.2 Future Work

- Currently, the attribute detectors used in Chapter 3 are sensitive to class imbalance – which is an inherent risk in attribute training. The framework in that chapter is eventually made robust to individual classifier error (after attribute-weighting) and we also find that this effect is less pronounced on different classifiers (by switching to LDA classifiers in Chapter 4), however since overall discriminative performance improves as a function of the average accuracy of all attribute classifiers, solving the imbalance problem remains a worthwhile objective. SVMs operate well on balanced data, but with imbalanced data tend toward predicting the majority class since the separating hyperplane becomes skewed toward the minority class, resulting in abnormally high false negative predictions [12]. Experiments were performed in order to quantify whether standard solutions such as oversampling the minority class or synthesising new instances [29] could alleviate this problem when training the SVM classifiers but did not prove to be helpful except for majority-class subsampling as detailed in Chapter 3. Several further options exist that deserve attention, such as (i) acquiring more data and labels, or (ii) adoption of a more interpretable classification model such as Decision Trees as applied by Liu *et al*. in [116], which provide both a human-readable solution that may inspire a more effective classification model, as well as simultaneously being more robust toward the class imbalance problem by incorporating a measure of class proportion to augment the standard metric, information gain.

- In Chapter 4, although we successfully map the Internet attributes to the expert ontology created in Chapter 3, (i) the Internet attributes themselves are not immediately as directly interpretable by humans and thus do not facilitate zero-shot re-identification (and therefore

retrieval queries). Furthermore, (ii) the meta-text upon which the clustering step operates is inherently noisy due to being unconstrained and unfiltered beyond standard natural language processing methods such as removing "stop" words, and because there is no guarantee the meta-text refers to the appearance of humans detected within the corresponding photograph. To address (i), a promising direction would be to again exploit the available labels produced in Chapter 3 in order to investigate the possibility of employing self-training, a bootstrap technique, using seed images of labelled attribute detections from VIPeR and other available datasets in conjunction with the large volume of already acquired Internet data. Self-training begins with an initial model trained on fully labelled data, and then used to estimate labels on a pool of data where the labels are unknown. A proportion of these estimated labels are added into the training pool, and the model is expected to improve after each subsequent iteration. Since the Internet, over time, continually makes new photographs available it is expected that the system could therefore continue to improve *ad infinitum* in a similar fashion to [30], particularly if the meta-data where incorporated as a prior during instance selection, addressing (ii). This strategy could alleviate both the class-imbalance problem discussed earlier, as well as providing a directly human-interprettable mid-level semantic representation from Internet data.

- There are several open issues for expanding Chapter 5's transfer-learning framework in order to improve performance and further reduce the amount of required annotation for good performance on unseen camera-pair domains. So far we have only used simplistic colour features and absolute performance should improve using better features as input. Additionally, multiple features can readily be included in our MKL framework, as well as the ability to fully incorporate fusion between LLFs and attributes – this is a crucial area of investigation since LLFs and attributes are diverse and complementary cues for re-identification. Another crucial aspect is the ability to transfer attribute classifiers between individual camera domains in order to avoid per-camera annotation cost. With regards to negative instance selection, we thus far randomly selected 10 negative pairs per positive pair for training although we note Re-identification accuracy can be increased at the cost of additional computation by increasing this ratio [6]. More interestingly we believe active learning or instance mining approaches to optimally select the right instances from the quadratic number of pairs is an important open question. Finally, we could also transduc-

tively exploit the unlabelled data distribution in the target domain, and eventually move towards completely annotation free transfer learning for re-identification.

- Chapter 6 introduces a provocation to the field of re-identification: we posit that the majority of re-identification work to date is unable to function for views that exhibit continuous view transformation and investigate several new variations on the standard re-identification paradigm. We enjoyed some success in our approach, providing an initial algorithm for the new paradigm via an unsupervised domain adaptation method that improved parity between disjoint "flights" (domains) by aligning the feature distributions. While re-identification performance in the "within-flight" case appears to improve following the application of domain adaptation this is likely related to how much motion is present in the entire flight *i.e*. if the UAV is relatively stable throughout then domain adaptation helps uniformly throughout the flight. However, if the surveillance video undergoes more dramatic view transformations such as those caused by more radical manoeuvring by the UAV, we expect this advantage to be much less. Therefore, for more robust re-identification during these cases it would be worthwhile to investigate more comprehensive solutions to this problem. Several possibilities for research in this direction exist. A simple extension might involve learning disjoint models for re-identification using human detections featuring motion-blur in a particular direction and dynamically switching to the relevant model depending on the present orientation and velocity of the UAV. A more generalisable approach would be to apply an online, unsupervised domain-adaptation algorithm across a temporal sliding-window of detections in order to "smooth" the distribution change between blocks of consecutive frames; the assumption being that multiple detections of the same person will be temporally proximal and thus online domain adaptation will facilitate the reconciliation of these detections into a single identity.

Bibliography

[1] The Picture Is Not Clear: How many CCTV surveillance cameras in the UK. Technical report, British Security Industry Association, London, England, 2013.

[2] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *European Conference on Machine Learning*, pages 39–50, 2004.

[3] A. Albiol, J. Oliver, and J. M. Mossi. Who is who at different cameras: people re-identification using depth cameras. *IET Computer Vision*, 6(5):378, 2012.

[4] L. An, M. Kafai, S. Yang, and B. Bhanu. Reference-Based Person Re-Identification. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2013.

[5] R. Armitage. To CCTV or not to CCTV? A review of current research into the effectiveness of CCTV systems in reducing crime. Technical Report 226171, Nacro, London, England, 2002.

[6] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning Implicit Transfer for Person Re-identification. In *European Conference on Computer Vision, International Workshop on Re-identification*, pages 381–390, Florence, Italy, 2012.

[7] T. Avraham and M. Lindenbaum. Learning Appearance Transfer for Person Re-identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*. Springer London, London, England, 2014.

[8] F. R. Bach, G. R. G. Lanckreit, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *International Conference on Machine Learning*, 2004.

[9] D. Baltieri, R. Vezzani, and R. Cucchiara. SARC3D: a new 3D body model for people tracking and re-identification. pages 197–206, Berlin, Heidelberg, 2011. Springer-Verlag.

[10] D. Baltieri, R. Vezzani, and R. Cucchiara. Learning Articulated Body Models for People Re-identification. In *ACM International Conference on Multimedia*, pages 557–560, 2013.

[11] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. Re-identification with RGB-D Sensors. In *European Conference on Computer Vision, International Workshop on Re-identification*, pages 433–442, Florence, Italy, 2012.

[12] R. Batuwita and V. Palade. Class Imbalance Learning Methods For Support Vector Machines. In H. He and Y. Ma, editors, *Imbalanced Learning: Foundations, Algorithms and Applications*. John Wiley & Sons, Inc., 2012.

[13] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision*, volume 3951, pages 404–417, Graz, Austria, 2006. Springer.

[14] L. Bazzani, M. Cristani, and V. Murino. SDALF: Modeling Human Appearance with Symmetry-Driven Accumulation Of Local Features. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*. Springer London, London, England, 2014.

[15] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot Person Re-identification by HPE signature. In *International Conference on Pattern Recognition*, pages 1413–1416, 2010.

[16] J. Beis and D. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1000–1006, 1997.

[17] G. Berdugo, O. Soceanu, Y. Moshe, D. Rudoy, and I. Dvir. Object Reidentification In Real World Scenarios Across Multiple Non-overlapping Cameras. *Signal Processing*, pages 1806–1810, 2010.

[18] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676, Heraklion, Crete, Greece, 2010. Springer-Verlag.

[19] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, 1969.

[20] A. Bialkowski, P. J. Lucey, W. Xinyu, and S. Sridharan. Person Re-Identification Using Group Information. *Digital Image Computing: Techniques and Applications*, 2013.

[21] S. Bąk, E. Corvée, F. Brémond, and M. Thonnat. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 435–440, Aug. 2010.

[22] S. Bąk, E. Corvée, F. Brémond, and M. Thonnat. Multiple-shot human re-identification by Mean Riemannian Covariance Grid. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 179–184. IEEE, 2011.

[23] S. Bąk, S. Zaidenberg, B. Boulay, and F. Brémond. Improving Person Re-identification by Viewpoint Cues. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2014.

[24] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. D. Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[25] Y. Brand, T. Avraham, and M. Lindenbaum. Transitive Re-identification. *British Machine Vision Conference*, (3):46.1–46.11, 2013.

[26] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[27] B. Cancella and T. M. Hospedales. Open-World Person Re-Identification by Multi-Label Assignment Inference. In *British Machine Vision Conference*, 2014.

[28] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1 —- 27:27, 2011.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June 2002.

[30] X. Chen, A. Shrivastava, and A. Gupta. NEIL : Extracting Visual Knowledge from Web Data. In *IEEE International Conference on Computer Vision*, 2013.

[31] D. S. Cheng and M. Cristani. Person Re-identification by Articulated Appearance Matching. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 139–160. Springer, London, 2014.

[32] D. S. Cheng, M. Cristani, V. Morego, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference*, 2011.

[33] E. D. Cheng and M. Piccardi. Matching of Objects Moving Across Disjoint Cameras. In *IEEE International Conference on Image Processing*, 2006.

[34] R. Clarke. Understanding the drone epidemic. *Computer Law & Security Review*, 30(3):230–246, June 2014.

[35] N. Dadashi, A. W. Stedmon, and T. P. Pridmore. Semi-automated CCTV surveillance: the effects of system confidence, system accuracy and task complexity on operator vigilance, reliance and workload. *Applied Ergonomics*, 44(5):730–738, 2013.

[36] W. Dai, Q. Yang, G. G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *International Conference on Machine Learning*, pages 193–200, New York, NY, USA, 2007. ACM.

[37] A. Dantcheva and J. Dugelay. Frontal-to-side face re-identification based on hair skin and clothes patches. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 309–313. IEEE, 2011.

[38] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, Oct. 2010.

[39] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-Theoretic Metric Learning. In *International Conference on Machine Learning*, pages 209–216, Corvallis, OR, 2007.

[40] I. O. de Oliveira and J. L. d. S. Pio. People Reidentification in a Camera Network. In *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 461–466, Dec. 2009.

[41] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Asian Conference on Computer Vision*, pages 501–512, Berlin, Heidelberg, 2011. Springer-Verlag.

[42] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daume, A. C. Berg, and T. L. Berg. Detecting visual text. In *North American Chapter of the Association for Computational Linguistics*, pages 762–772, 2012.

[43] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2014.

[44] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–61, Apr. 2012.

[45] L. Duan, I. I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer SVM for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[46] F. Orabona, L. Jie, and F. Orabona. Ultra-fast optimization algorithm for sparse multi kernel learning. In *International Conference on Machine Learning*, 2011.

[47] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2367, June 2010.

[48] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2352–2359, 2010.

[49] L. Fei-Fei, R. Fergus, and P. Perona. One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[50] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–45, Sept. 2010.

[51] V. Ferrari and A. Zisserman. Learning visual attributes. *Neural Information Processing Systems*, 2007.

[52] D. Figueira, L. Bazzani, M. Cristani, A. Bernardino, V. Murino, and I. Italiano. Semi-supervised Multi-feature Learning for Person Re-identification. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 111–116, 2013.

[53] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised Multi-feature Learning for Person Re-identification. *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2013.

[54] R. L. Finn and D. Wright. Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications. *Computer Law & Security Review*, 28(2):184–194, Apr. 2012.

[55] P. Földiák. Sparse coding in the primate cortex. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1064–1068. MIT Press, second edition, 2002.

[56] P.-E. Forssén. Maximally Stable Colour Regions for Recognition and Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[57] E. W. Frew, C. Dixon, J. Elston, and M. Stachura. Active sensing by unmanned aircraft systems in realistic communication environments. In *Proceedings of the International Federation of Automatic Control*, 2009.

[58] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute Learning for Understanding Unstructured Social Activity. In *European Conference on Computer Vision*, pages 530–543. Springer Berlin Heidelberg, Florence, Italy, 2012.

[59] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):303–16, Feb. 2014.

[60] G. Gerrard, G. Parkins, I. Cunningham, W. Jones, S. Hill, and S. Douglas. National CCTV Strategy. Technical Report October, UK Home Office, London, England, 2007.

[61] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley. Person Reidentification Using Spatiotemporal Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, 2006.

[62] M. Gill and A. Spriggs. Assessing the impact of CCTV. Technical report, Home Office Research, Development and Statistics Directorate, London, England, 2005.

[63] M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, K. Deena, J. Kilworth, R. Little, and D. Swain. Control room operation: findings from control room observations. Technical report, UK Home Office, London, United Kingdom, May 2005.

[64] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, June 2012.

[65] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales. The Re-Identification Challenge. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 1–21. Springer London, London, England, 2014.

[66] S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-identification*. Springer, 2014.

[67] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, page 5, 2007.

[68] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, pages 262–275, Marseille, France, 2008.

[69] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric Learning Approaches for Face Identification. In *IEEE International Conference on Computer Vision*, 2009.

[70] M. Hahnel, D. Klunder, and K.-F. Kraiss. Color and texture features for person recognition. In *International Joint Conference on Neural Networks*, pages 647–652, 2004.

[71] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person Re-identification In Multi-camera System By Signature Based On Interest Point Descriptors Collected On Short Video Sequences. In *IEEE International Conference on Distributed Smart Cameras*, pages 1–6, 2008.

[72] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[73] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arxiv*, 2012.

[74] M. Hirzer, C. Beleznai, M. Köstinger, P. M. Roth, and H. Bischof. Dense Appearance

Modeling And Efficient Learning Of Camera Transitions For Person Re-identification. In *IEEE International Conference on Image Processing*, pages 1–4, 2012.

[75] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. *Scandinavian Conference on Image Analysis*, 2011.

[76] M. Hirzer, P. M. Roth, and H. Bischof. Person Re-identification by Efficient Impostor-Based Metric Learning. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 203–208, Sept. 2012.

[77] M. Hirzer, P. M. Roth, K. Martin, H. Bischof, and M. Köstinger. Relaxed Pairwise Learned Metric for Person Re-identification. In *European Conference on Computer Vision*, pages 780–793, Florence, Italy, 2012.

[78] T. M. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying Rare and Subtle Behaviours: A Weakly Supervised Joint Topic Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[79] C. J. Howard, T. Troscianko, I. D. Gilchrist, A. Behera, and D. C. Hogg. Searching for threat: factors determining performance during CCTV monitoring. In D. De Waard, J. Godthelp, F. L. Kooi, and K. A. Brookhuis, editors, *Human Factors, Security and Safety*. Shaker Publishing, 2009.

[80] Y. H. Hu and J.-N. Hwang, editors. *Handbook of neural network signal processing*. Taylor & Francis, 2001.

[81] C.-H. Huang, Y.-T. Wu, and M.-Y. Shih. Unsupervised pedestrian re-identification for loitering detection. *Advances in Image and Video Technology*, pages 771–783, 2009.

[82] A. K. Jain, S. C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication*, pages 731—-738, Hong Kong, 2004.

[83] L. Jie, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *IEEE International Conference on Computer Vision*, 2011.

[84] V. John, G. Englebienne, and B. Krose. Solving Person Re-identification in Non-overlapping Camera using Efficient Gibbs Sampling. In *British Machine Vision Conference*, 2013.

[85] I. T. Jolliffe. *Principle Component Analysis*. Springer, New York, New York, USA, second edi edition, 2002.

[86] K. Jüngling and M. Arens. Local Feature Based Person Reidentification in Infrared Image Sequences. *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 448–455, Aug. 2010.

[87] S. Karaman and A. D. Bagdanov. Identity Inference : Generalizing Person Re-identification Scenarios. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *European Conference on Computer Vision, International Workshop on Re-identification*, volume 7583, pages 443–452, Florence, Italy, 2012. Springer.

[88] S. Karaman, G. Lisanti, A. D. Bagdanov, and A. Delbimbo. From Re-identification to Identity Inference: Labeling Consistency by Local Similarity Constraints. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*. Springer, London, 2014.

[89] H. Keval. CCTV Control Room Collaboration and Communication: Does it Work? In *Proceedings of Human Centred Technology Workshop*, pages 11–12, 2006.

[90] H. Keval and M. A. Sasse. "Not the Usual Suspects": A Study of Factors Reducing the Effectiveness of CCTV. *Security Journal*, pages 1–21, Oct. 2008.

[91] M. I. Khedher, M. A. El Yacoubi, and B. Dorizzi. Multi-shot SURF-based Person Re-identification via sparse representation. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 159–164, 2013.

[92] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the Damage of Dataset Bias. In *European Conference on Computer Vision*, Florence, Italy, Oct. 2012.

[93] W. Kohler. *The task of gestalt psychology*. Princeton University Press, Princeton, 1969.

[94] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric

learning from equivalence constraints. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295, 2012.

[95] N. Kumar and P. Belhumeur. FaceTracer: A search engine for large collections of images with faces. In *European Conference on Computer Vision*, pages 1–14, Marseille, France, 2008.

[96] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[97] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[98] C.-H. Kuo, S. Khamis, and V. Shet. Person re-identification using semantic color names and RankBoost. In *Workshop on the Applications of Computer Vision*, pages 1–7, 2013.

[99] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[100] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–65, Mar. 2013.

[101] R. Layne, T. M. Hospedales, and S. Gong. Person Re-identification by Attributes. In *British Machine Vision Conference*, 2012.

[102] R. Layne, T. M. Hospedales, and S. Gong. Towards Person Identification and Re-identification with Attributes. In *European Conference on Computer Vision, International Workshop on Re-identification*, pages 402–412, Florence, Italy, 2012.

[103] R. Layne, T. M. Hospedales, and S. Gong. Domain Transfer for Person Re-identification. In *Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, Barcelona, Spain, 2013.

[104] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based Re-identification. In S. Gong,

M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 93–117. Springer London, London, England, 2014.

[105] A. Li, L. Liu, and S. Yan. Can Feature-Based Inductive Transfer Learning Help Person Re-identification? In *IEEE International Conference on Image Processing*, 2013.

[106] A. Li, L. Liu, and S. Yan. Clothes Attributes Assisted Person Re-identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 119–138. Springer, 2014.

[107] W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[108] W. Li, Y. Wu, M. Mukunoki, and M. Minoh. Coupled metric learning for single-shot versus single-shot person reidentification. *Proceedings of SPIE*, 52(2), 2013.

[109] W. Li, R. Zhao, and X. Wang. Human Reidentification with Transferred Metric Learning. In *Asian Conference on Computer Vision*, 2012.

[110] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2013.

[111] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer Learning by Borrowing Examples for Multiclass Object Detection. *Neural Information Processing Systems*, pages 1–9, 2011.

[112] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person Re-identification: What Features Are Important? In *European Conference on Computer Vision, International Workshop on Re-identification*, pages 391–401, Florence, Italy, 2012.

[113] C. Liu, S. Gong, C. C. Loy, and X. Lin. Evaluating Feature Importance for Re-identification. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 203–. Springer, London, 2014.

[114] D. Liu and J. Nocedal. On the limited memory method for large scale optimization. *Mathematical Programming B*, 45(3):503–528, 1989.

[115] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, 2011.

[116] W. Liu, S. Chawla, D. Cieslak, and N. Chawla. A Robust Decision Tree Algorithm for Imbalanced Data Sets. In *SIAM International Conference on Data Mining*, 2010.

[117] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, 45(12):4204–4213, 2012.

[118] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer Feature Learning with Joint Distribution Adaptation. In *IEEE International Conference on Computer Vision*, 2013.

[119] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004.

[120] C. C. Loy, T. Xiang, and S. Gong. Time-Delayed Correlation Analysis for Multi-Camera Activity Understanding. *International Journal of Computer Vision*, 90(1):106–129, May 2010.

[121] A. J. Ma, P. C. Yuen, and J. Li. Domain Transfer Support Vector Ranking for Person Re-Identification without Target Camera Label Information. In *IEEE International Conference on Computer Vision*, 2013.

[122] B. Ma, Y. Su, and F. Jurie. BiCov: a novel image representation for person re-identification and face verification. In *British Machine Vision Conference*, pages 57.1–57.11. British Machine Vision Association, 2012.

[123] B. Ma, Y. Su, and F. Jurie. Local Descriptors encoded by Fisher Vectors for Person Re-identification. In *European Conference on Computer Vision, International Workshop on Re-identification*, Florence, Italy, 2012.

[124] D. J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 4th edition, 2003.

[125] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18(3-4):233–247, Mar. 2007.

[126] L. Marchesotti and F. Perronnin. Learning beautiful (and ugly) attributes. In *British Machine Vision Conference*, Bristol, England, 2013.

[127] L. Mejías, J. F. Correa, I. Mondrag, and P. Campoy. COLIBRI: A vision-Guided UAV for Surveillance and Visual Inspection. In *International Conference on Robotics and Automation*, pages 2760–2761, Rome, Italy, 2007.

[128] A. Mignon and F. Jurie. PCCA: A New Approach for Distance Learning from Sparse Pairwise Constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[129] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern Recognition*, 2004.

[130] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering Analysis and an algorithm. *Neural Information Processing Systems*, 2001.

[131] J. Nocedal and S. Wright. *Numerical optimization*. Springer-Verlag, 2nd edition, 2006.

[132] T. Nortcliffe. *People Analysis CCTV Investigator Handbook*. Home Office Centre of Applied Science and Technology, London, England, 2011.

[133] N. F. Noy and D. L. Mcguinness. Ontology Development 101 : A Guide to Creating Your First Ontology. *Development*, 32(1):1–25, 2000.

[134] L. Ogden. Drone Ecology. *BioScience*, 63(9):776–776, Sept. 2013.

[135] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation Of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[136] F. Orabona. DOGMA: a MATLAB toolbox for Online Learning, 2009.

[137] F. Orabona, L. Jie, and B. Caputo. Online-batch strongly convex Multi Kernel Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–794, June 2010.

[138] J. Orwell, P. Remagnino, and G. A. Jones. Multiple camera color tracking. In *IEEE Workshop on Visual Surveillance*, 1999.

[139] D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991.

[140] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Data and Knowledge Engineering*, 22(10):1345–1359, Oct. 2010.

[141] D. Parikh and K. Grauman. Relative Attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 503–510, 2011.

[142] U. Park, A. Jain, I. Kitahara, and N. Hagita. ViSE: Visual Search Engine Using Multiple Networked Cameras. In *International Conference on Pattern Recognition*, pages 1204–1207, 2006.

[143] K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895.

[144] J. C. Platt. *Probabilities for SV Machines*, pages 61–74. MIT Press, 1999.

[145] F. Porikli. Inter-Camera Color Calibration by Correlation Model Function. In *IEEE International Conference on Image Processing*, 2004.

[146] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *British Machine Vision Conference*, 2008.

[147] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference*, pages 21.1–21.11, 2010.

[148] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative Multi-Label Video Annotation. In *ACM International Conference on Multimedia*, 2007.

[149] Y. Raja and S. Gong. Scalable Multi-camera Tracking in a Metropolis. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 417–441. Springer London, London, England, 2014.

[150] U. Ruckert and S. Kramer. Kernel-Based Inductive Transfer. In *European Conference on Machine Learning*, pages 220–233, 2008.

[151] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, 2010.

[152] P. Salvagnini, L. Bazzani, M. Cristani, and V. Murino. Person Re-identification with a PTZ camera: An Introductory Study. In *IEEE International Conference on Image Processing*, pages 3552–3556, Sept. 2013.

[153] R. Satta, G. Fumera, and F. Roli. A General Method for Appearance-Based People Search Based on Textual Queries. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *European Conference on Computer Vision, International Workshop on Re-identification*, volume 7583, pages 453–461, Florence, Italy, 2012. Springer.

[154] R. Satta, F. Pala, G. Fumera, and F. Roli. People Search with Textual Queries About Clothing Appearance Attributes. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*. Springer, London, England, 2014.

[155] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, MA, 2002.

[156] W. R. Schwartz. Scalable People Re-identification Based On A One-against-some Classification Scheme. In *IEEE International Conference on Image Processing*, 2012.

[157] W. R. Schwartz and L. S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. *Computer Graphics and Image Processing (SIBGRAPI)*, pages 322–329, Oct. 2009.

[158] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. Davis. Human detection using partial least squares analysis. In *IEEE International Conference on Computer Vision*, pages 24–31. IEEE, 2009.

[159] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[160] G. J. D. Smith. Behind the screens: Examining constructions of deviance and informal practices among CCTV control room operators in the UK. *Surveillance and Society*, 2:376–395, 2004.

[161] P. Smyth. Bounds on the mean classification error rate of multiple experts. *Pattern Recognition Letters*, 17:1253–1257, 1996.

[162] E. Sommerlade and I. Reid. Information-theoretic active scene exploration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[163] H. Su, J. Deng, and L. Fei-Fei. Crowdsourcing annotations for visual object detection. In

Workshop on Human Computation, Association for the Advancement of Artificial Intelligence, pages 40–46, 2012.

[164] R. Szeliski. *Computer Vision: Algorithms and Applications*, volume 5. Springer-Verlag New York Inc, 2010.

[165] S. Tahir and A. Cavallaro. Cost-effective features for re-identification in camera networks. *IEEE Transactions on Circuits and Systems for Video Technology*, X(c):1–1, 2014.

[166] J. Tang, X.-S. Hua, M. Wang, Z. Gu, G.-J. Qi, and X. Wu. Correlative Linear Neighborhood Propagation for Video Annotation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(2):409–416, 2009.

[167] R. Tarjan. Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing*, 1(2):146–160, June 1972.

[168] A. H. Tickner and E. C. Poulton. Monitoring up to 16 synthetic television pictures showing a great deal of movement. *Ergonomics*, 16:381–401, 1973.

[169] T. Tommasi, F. Orabona, and B. Caputo. Learning Categories from Few Examples with Multi Model Knowledge Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):928–941, Oct. 2013.

[170] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[171] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, Jan. 2012.

[172] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, 2006.

[173] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–23, July 2009.

[174] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *Workshop on the Applications of Computer Vision*, pages 1–8, Snowbird, Utah, 2009.

[175] E. Wallace and C. Diffley. CCTV control room ergonomics. Technical report, Police Scientific Development Branch of the Home Office, Hertfordshire, England, 1998.

[176] C. V. D. Walt and E. Barnard. Data characteristics that determine classifier performance. In *Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 160–165, 2006.

[177] L. Wang, F. Su, H. Zhu, and L. Shen. Active sensing based cooperative target tracking using UAVs in an urban area. *International Conference on Advanced Computer Control*, pages 486–491, 2010.

[178] X. Wang and R. Zhao. Person Re-identification: System Design and Evaluation Overview. In S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors, *Person Re-identification*, pages 350–370. Springer London, London, England, 2014.

[179] K. Q. Weinberger and L. K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 10:207–244, June 2009.

[180] C. Wilkinson and R. Evans. Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science & justice : journal of the Forensic Science Society*, 49(3):191–6, Sept. 2009.

[181] D. Williams. Effective CCTV and the challenge of constructing legitimate suspicion using remote visual images. *Journal of Investigative Psychology and Offender Profiling*, 4(2):97–107, 2007.

[182] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human Re-identification by Matching Compositional Template with Cluster Sampling. In *IEEE International Conference on Computer Vision*, number 1, 2013.

[183] K. Yanai and K. Barnard. Image region entropy: a measure of visualness of web images associated with one concept. In *ACM International Conference on Multimedia*, 2005.

[184] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient Color Names for Person Re-identification. In *European Conference on Computer Vision*, pages 536–551, 2014.

[185] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *Neural Information Processing Systems*, 17:1601–1608, 2004.

[186] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. In *IEEE International Conference on Computer Vision*, pages 3586–3593, June 2013.

[187] W.-S. Zheng, S. Gong, and T. Xiang. Associating Groups of People. In *British Machine Vision Conference*, 2009.

[188] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 649–656. IEEE, 2011.

[189] W.-S. Zheng, S. Gong, and T. Xiang. Quantifying and Transferring Contextual Information in Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(8):1–14, Aug. 2011.

[190] W.-S. Zheng, S. Gong, and T. Xiang. Unsupervised selective transfer learning for object recognition. In *Asian Conference on Computer Vision*, pages 527–541. Springer, 2011.

[191] W.-S. Zheng, S. Gong, and T. Xiang. Transfer Re-identification: From Person to Set-based Verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, USA, 2012.

[192] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013.

[193] J. Zhu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Pedestrian Attribute Classification in Surveillance : Database and Evaluation. In *ICCV workshop on Large-Scale Video Search and Mining*, 2013.

[194] X. Zhu and X. Wu. Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. *Artificial Intelligence Review*, 22(1):177–210, 2004.